Multi Protocol Label Switching (MPLS) and L2/L3 VPNs

IERG5090

Spring, 2017

Wing C. Lau

Acknowledgements

- Many of the slides used in this chapter are adapted from the following sources:

  - John Drake, Walt Wimer, Fore Systems.
  - Tim G. Griffin, Cambridge University
  - Peter Tomsu, Cisco
  - Jeff Doyle, Jeff Doyle and Associates Inc.
  - Li Yi, UC Berkeley
  - Jurrie Vandenbreekel, Spirent Communications
  - Pramoda Nallur, Alcatel-Lucent
  - Jim Kurose and Keith Ross, "Computer Networks – A top-down approach " 6th Ed., published by Addison Wesley.
  - Yaakov J. Stein, "VPLS", RAD Data communications.
  - Ferit Yegenoglu, "Introduction to MPLS-based VPNs", ISOCORE.
  - Bruno De Troch, "VPLS", Juniper Networks.

Recap: What kind of traffic engineering can be done with existing IGPs

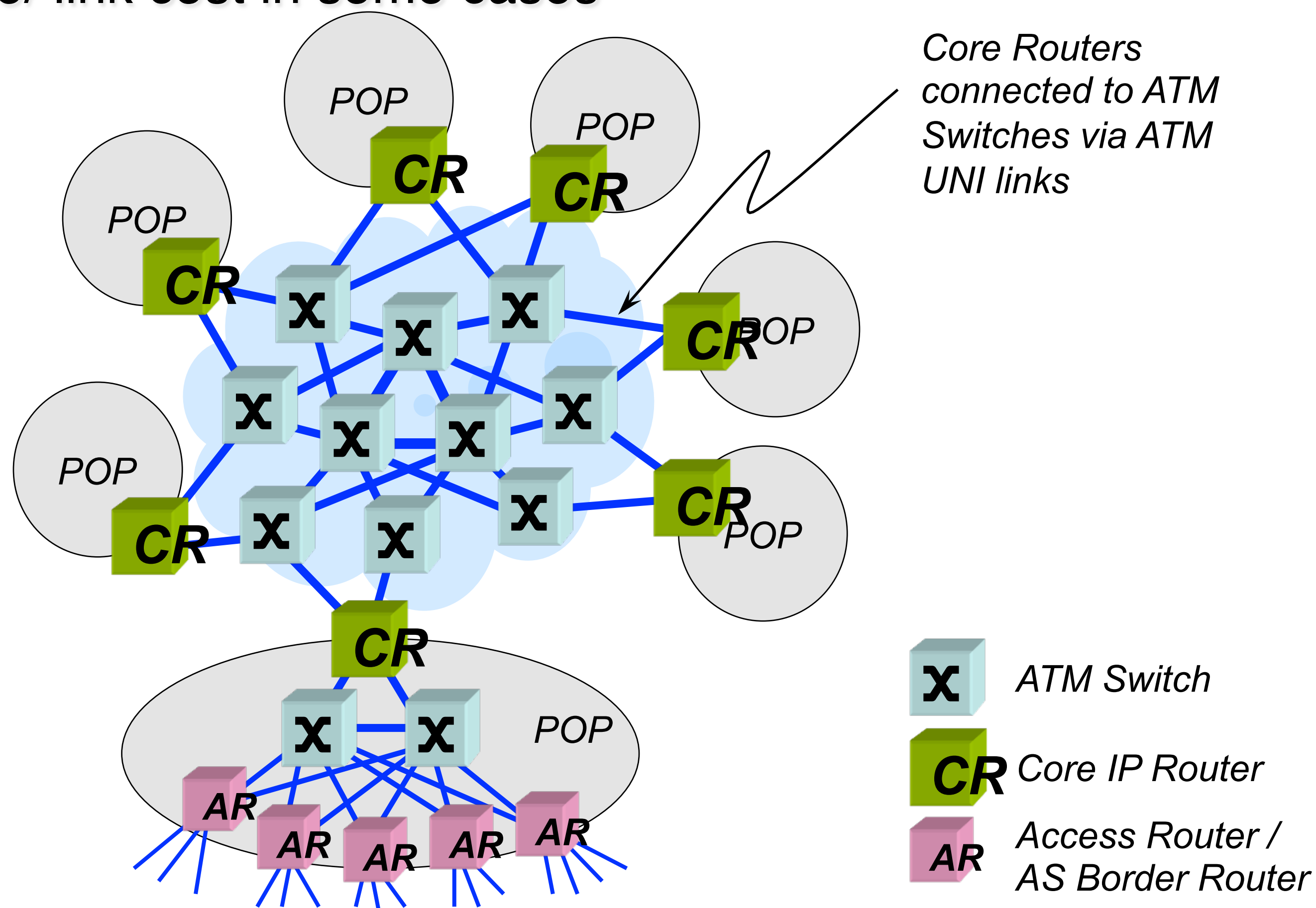
- On Intradomain routing:
  - tune link-metrics used for Shortest path computation
  - set link to default values, usually inversely proportional to link-speed, static weight (i.e. no change except link failure)
  - dynamic link metrics, e.g. load-dependent (EIGRP), can be dangerous
  - Equal Cost Multiple Path (ECMP) routing to give more flexibility to do load sharing across multiple shortest paths
  - depart from shortest-path routing can lead to routing loops if not careful
  - Hard to find (NP-hard) the required link-weights in order to realize a given routing pattern.

Asynchronous Transfer Mode: ATM

- **1990's -2000 standard for high-speed** (155Mbps to 622 Mbps and higher) *Broadband Integrated Service Digital Network* architecture
- Goal: *integrated, end-to-end transport for carrying voice, video, data*
  - meeting timing/QoS requirements of voice, video (versus Internet best-effort model)
  - "next generation" telephony: technical roots in telephone world
  - packet-switching (fixed length packets, called "cells") using virtual circuits

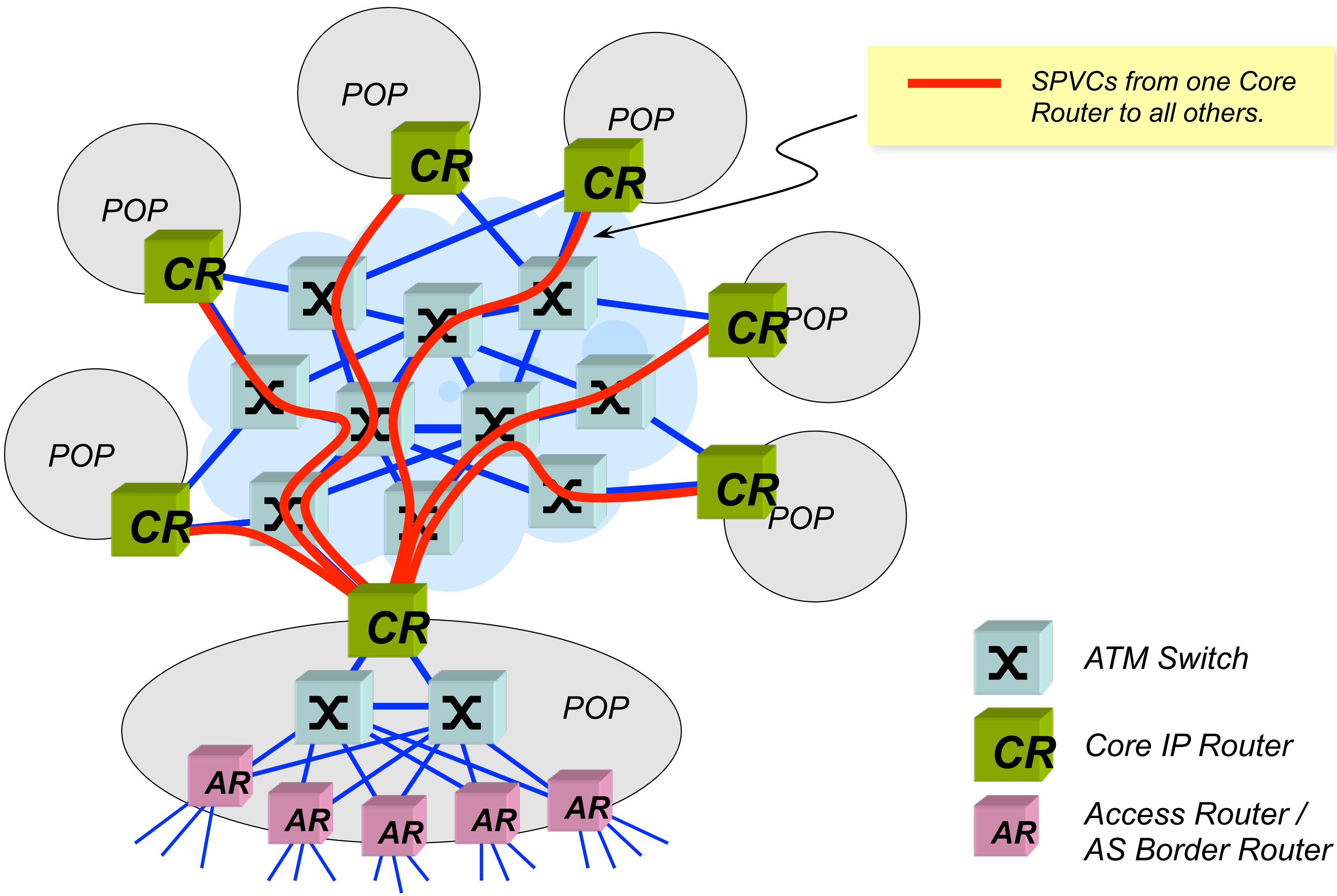
Common Traffic Engineering practice in IP networks

- For Intradomain routing: Before MPLS, most big ISPs implement the IP-over-ATM model, many already migrated to MPLS:
  - Use an ATM cloud with Permanent Virtual Circuits (PVCs) to provide DIRECT connection between each router-pair => facilitate bandwidth management and route predictability ; may save some interface/ link cost in some cases



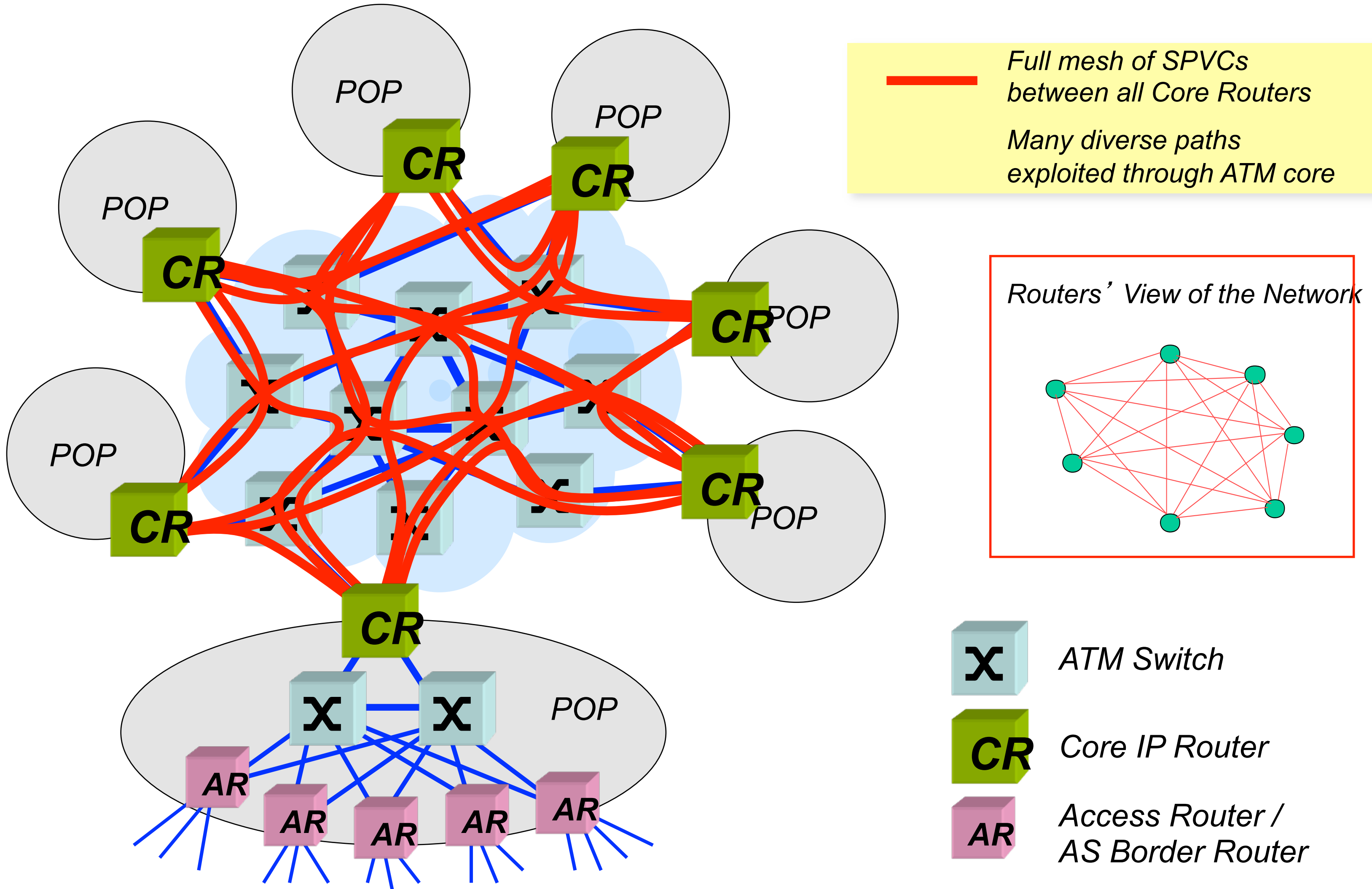
Common IP Traffic Engineering in practice (cont’d)

- Full-mesh Layer-3, i.e. router, peering is required => IGP scalability problem

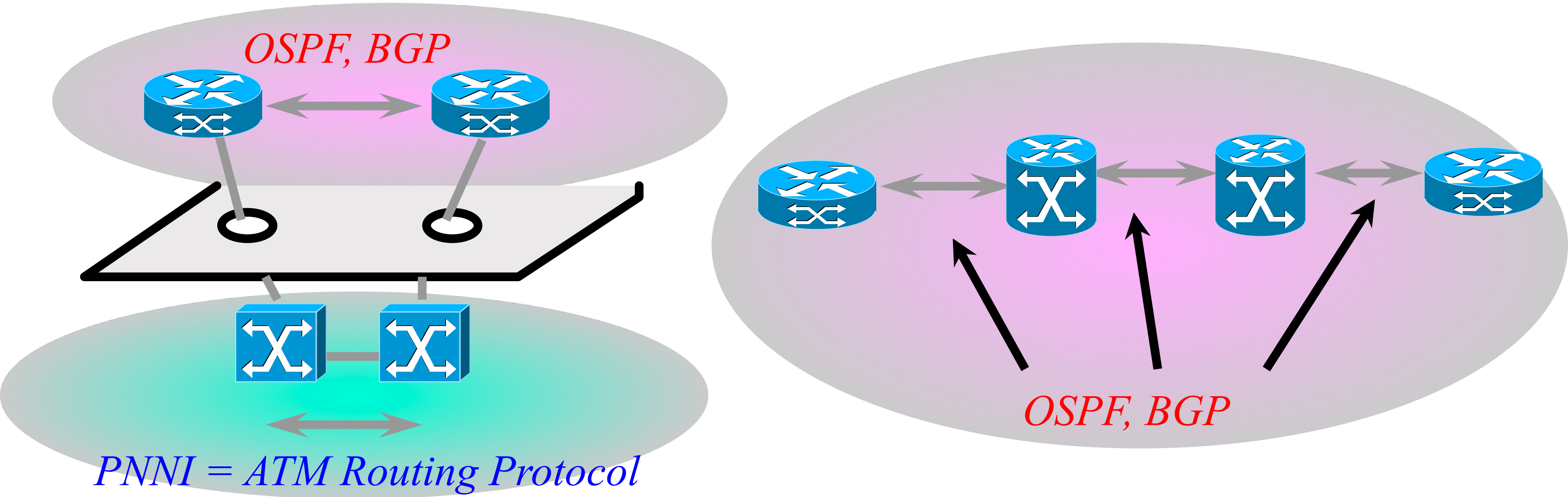


Common IP Traffic Engineering in practice (cont’d)

- Full-mesh Layer-3, i.e. router, peering is required => IGP scalability problem
- The Overlay model => High cost for maintaining 2 separate networks: one ATM, one IP ; Many ISPs have used MPLS to replace ATM’s role here.



IP-over-ATM Overlay Model vs. MPLS Peer Model



IP-over-ATM Overlay Model

Routers and Switches totally isolated
Routers have no idea of ATM Topology
IP features must be approximately mapped into ATM

MPLS Peer Model

Routers and Switches totally integrated
Routers & Switches share topology
IP features directly supported by the MPLS switches

MPLS vs. ATM

- Many basic MPLS concepts borrowed from ATM:

	ATM	MPLS
Switching Field	VP / VC	Label (stackable)
Routable Objects	Virtual Circuits	Label Switched Paths (LSPs)
Source Routing	Designated Transit List	Explicit Route
Path Setup	PNNI Signaling	LDP, Modified/extended versions RSVP, BGP, OSPF, IS-IS

- ***To meet QoS requirements, even non-ATM LSRs will end up strongly resembling ATM switches:***

	ATM	MPLS
Queuing	Per-VC queuing	Per-LSP queuing
Traffic Scheduling	Weighted per-VC scheduling	Weighted per-LSP scheduling
QoS Routing	PNNI routing	RSVP-TE, CR-LDP (Constraint-based Routing LDP)

MPLS – Multi Protocol Label Switching

"The primary goal of the MPLS working group is to standardise a base technology that integrates the label swapping forwarding paradigm with network layer routing.

Label Swapping is expected to improve

- *price/performance of network layer routing*
- *scalability of the network layer*
- *provide greater flexibility in the delivery of (new) routing services*
  - *new routing services can be added without changing the forwarding paradigm*

RFC3031, Jan 2001.

MPLS Basic Terminology

- Label
  - A fixed-length (20-bit) header field to identify packets belonging to "virtual circuit", i.e. stream of packets
  - Local significance (link scope)
- Label Switched Paths (LSPs)
  - An MPLS virtual circuit
  - LSPs are unidirectional
- Label Switching Routers (LSRs)
  - Any router capable of supporting MPLS
- Forwarding Equivalence Classes (FECs)
  - All packets:
    - To be forwarded out the same interface
    - With the same forwarding treatment (CoS)
    - To the same next hop

Core mechanisms of MPLS

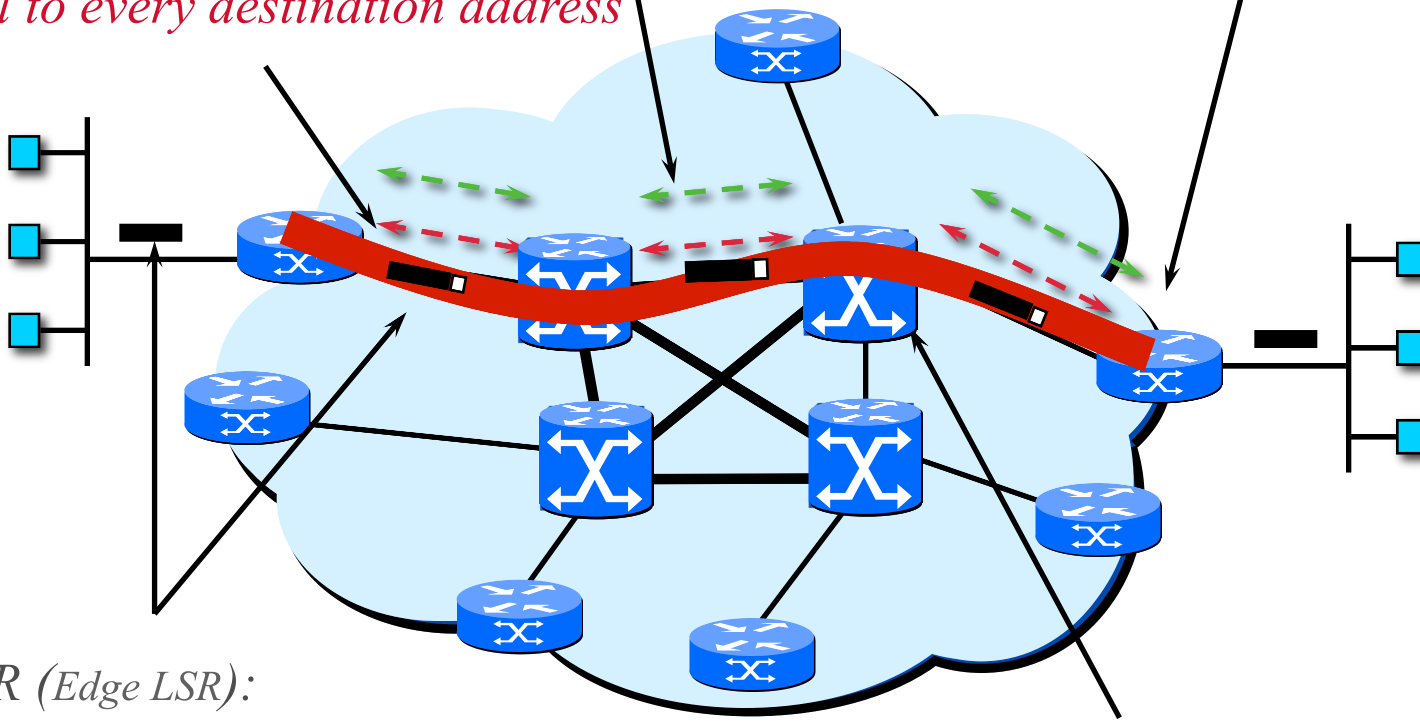
- **Semantics assigned to a stream label**
  - Labels are associated with specific streams of data.
- **Forwarding Methods**
  - Forwarding is simplified by the use of the short fixed length labels to identify streams.
  - Forwarding may require simple functions such as looking up a label in a table, swapping labels, and possibly decrementing and checking a TTL.
  - In some case MPLS may direct uses of underlying layer 2 forwarding.
- **Label Distribution Methods**
  - Allow nodes to determine which labels to use for specific streams.
  - This may use some sort of control exchange, and/or be piggybacked on a routing protocol.

MPLS Operations

1a. The Routed protocol (OSPF, IGRP,...) computes the shortest path to destination within the core

4.The last MPLS router removes label

1b. Some Label Distribution Protocol(e.g. LDP, RSVP-TE, MP-BGP) binds a label to every destination address



2. ELSR (Edge LSR):

- *Inbound router receives packets*
- *runs usual L3 services*
- *adds labels to packets*

3.LSR: Label Switch Router

- *switches packet based on label - Label Swapping*

Label-Switched Routers (LSR)s

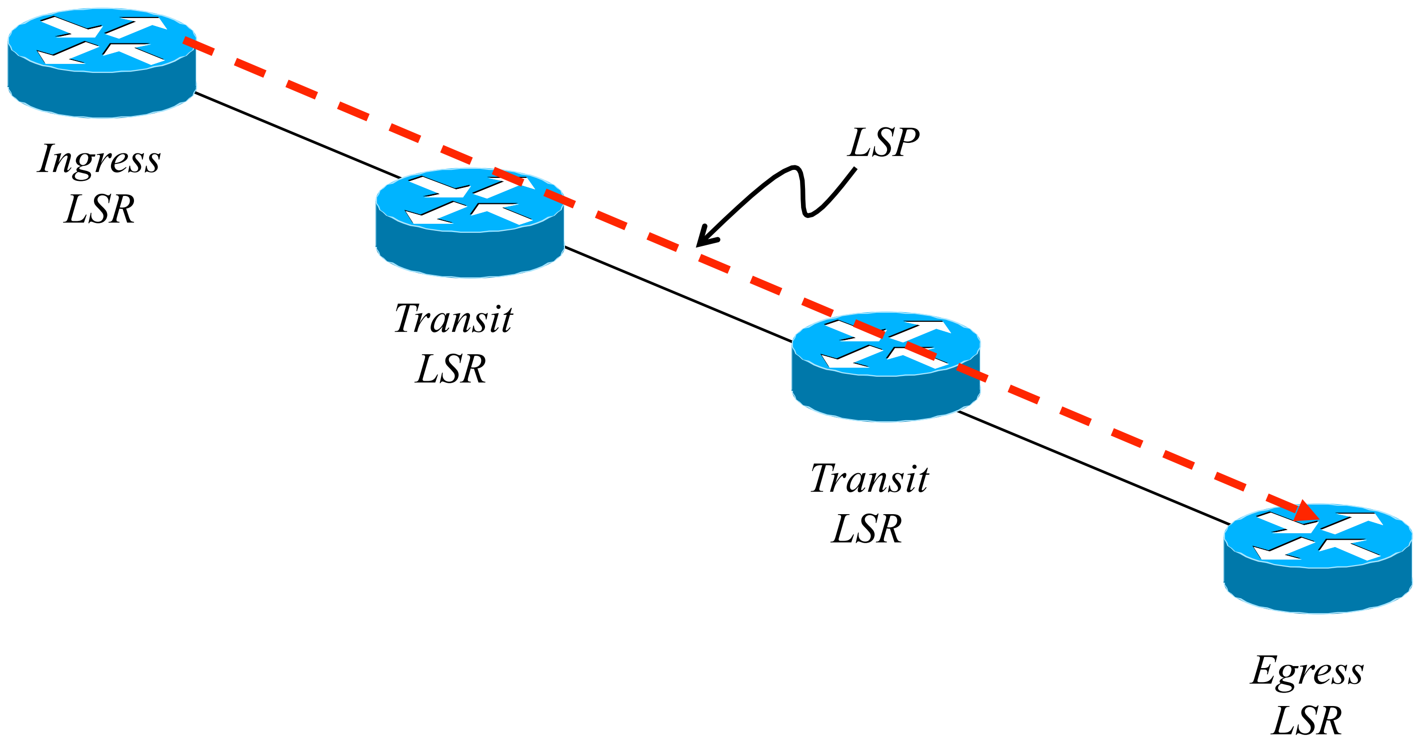
- Forwards packets to outgoing interface based on label value (don’t inspect IP address except the Edge-LSRs)
  - MPLS forwarding table distinct from IP forwarding tables
- signaling protocol needed to set up the labels
  - e.g. LDP (Label Distribution Protocol), or using extensions of BGP (MP-BGP), RSVP (RSVP-TE)
  - Forwarding possible along paths that IP alone would not allow (e.g., source-specific routing) !!

  => Facilitate the use of MPLS for traffic engineering
- CAN co-exist with IP-only routers

Forwarding Equivalent Class

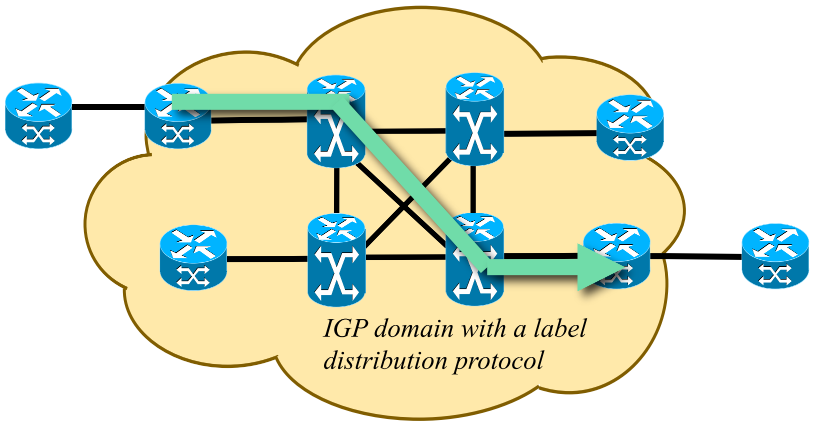
- IP Packets are classified into Forwarding Equivalent Class (FECs)
  - group of packets forwarded in the same manner, over the same path, with the same forwarding treatment
    - determined (by default) through the output of the IGP (or static routing)
  - each FEC corresponds to an IP destination prefix
    - destination-based unicast routing (default)
    - could be QOS, all BGP prefixes reachable via a particular exit point etc...

A Label Switched Path (LSP)

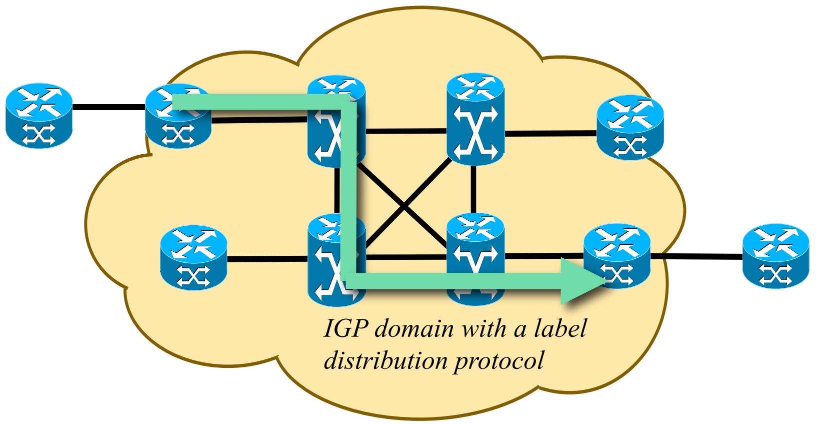


- *LSPs are unidirectional*
- *Ingress, transit, and egress are relative to a given LSP*
- *A given router can be ingress, egress, and transit for different LSPs*

Label Switched Path (LSP)



LSP follows IGP shortest path



LSP diverges from IGP shortest path

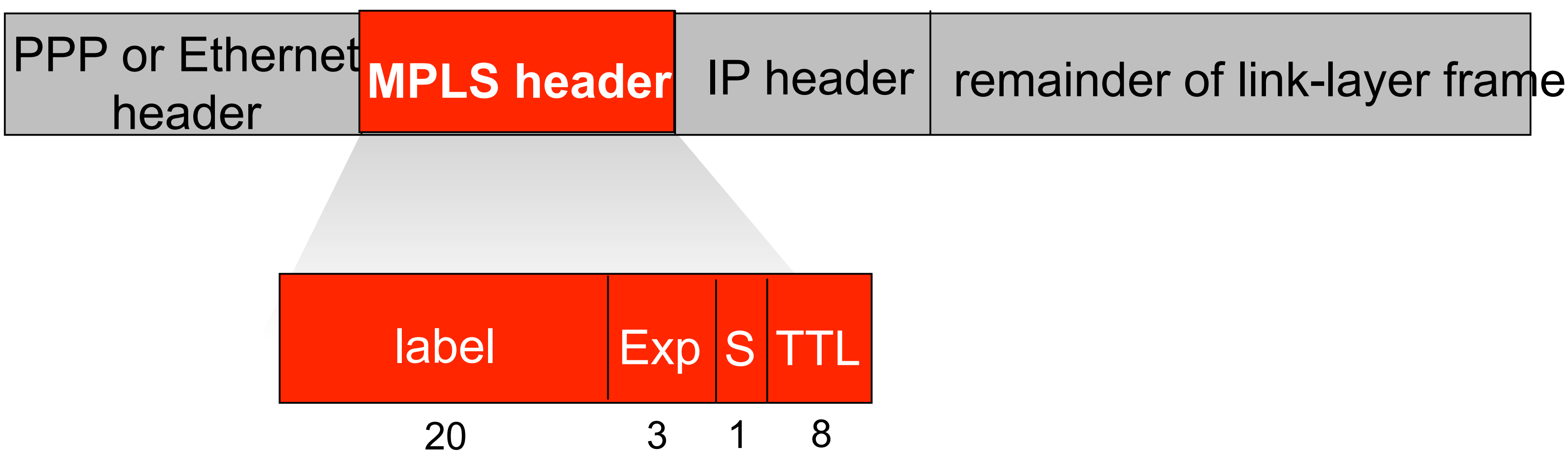
- **FEC is determined in LSR-ingress**
- **LSR-ingress to LSR-egress path is the same for packets of the same FEC**
- **LSPs are derived from IGP routing information**

- **LSPs may diverge from IGP shortest path**

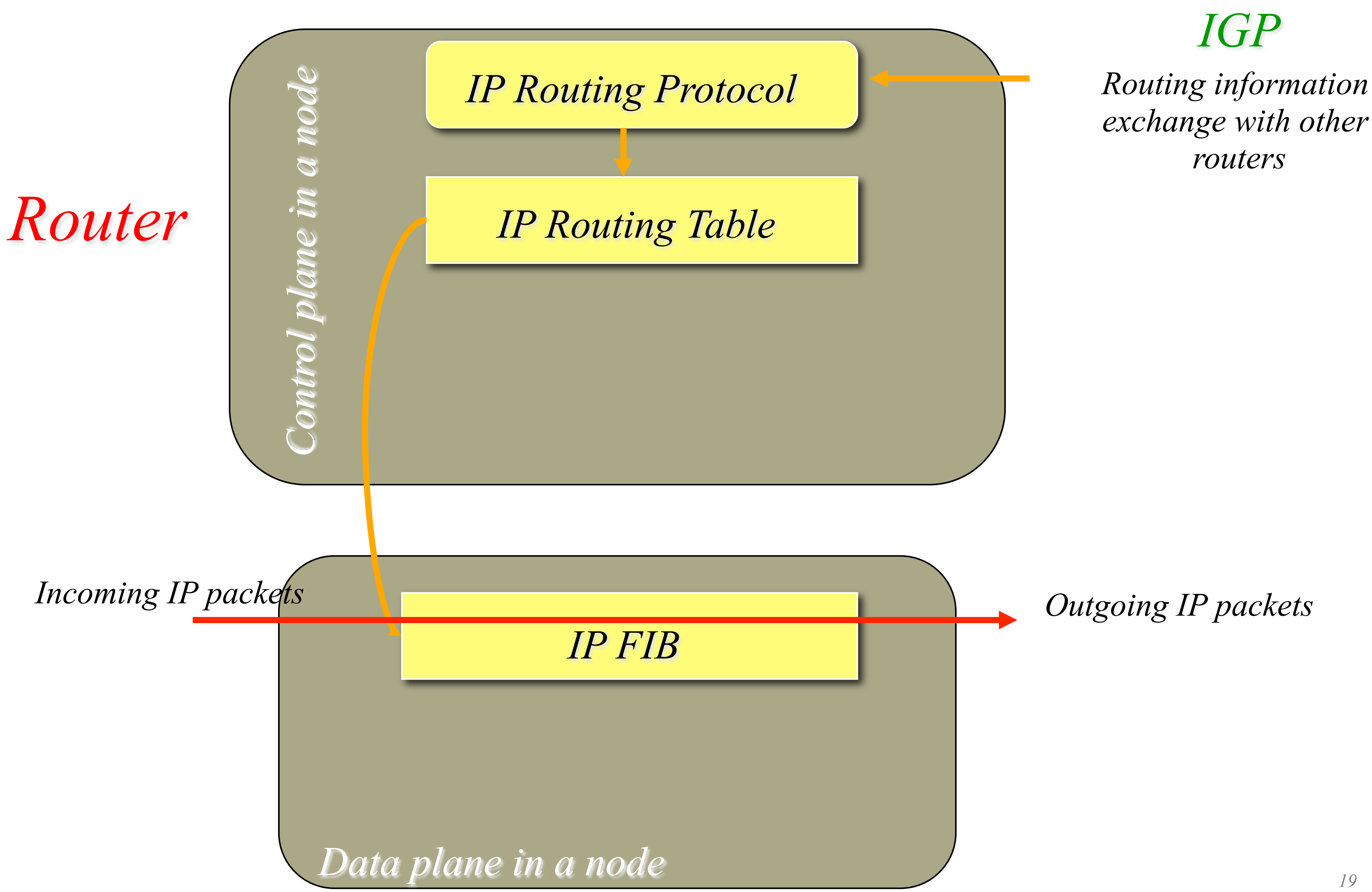
  LSP tunnels (explicit routing) with Traffic Engineering

Multiprotocol label switching (MPLS)

- initial goal: speed up IP forwarding by using fixed length label (instead of IP address) to do forwarding
  - borrowing ideas from Virtual Circuit (VC) approach
  - but IP datagram still keeps IP address!



Control-Plane to Data-Plane



MPLS Forwarding Component

Forwarding Component

also referred to as the data plane

responsible for forwarding packets/cells based on labels

uses a label forwarding database maintained by the label switch

Simple Label Swapping

MPLS Control Component

Control Component

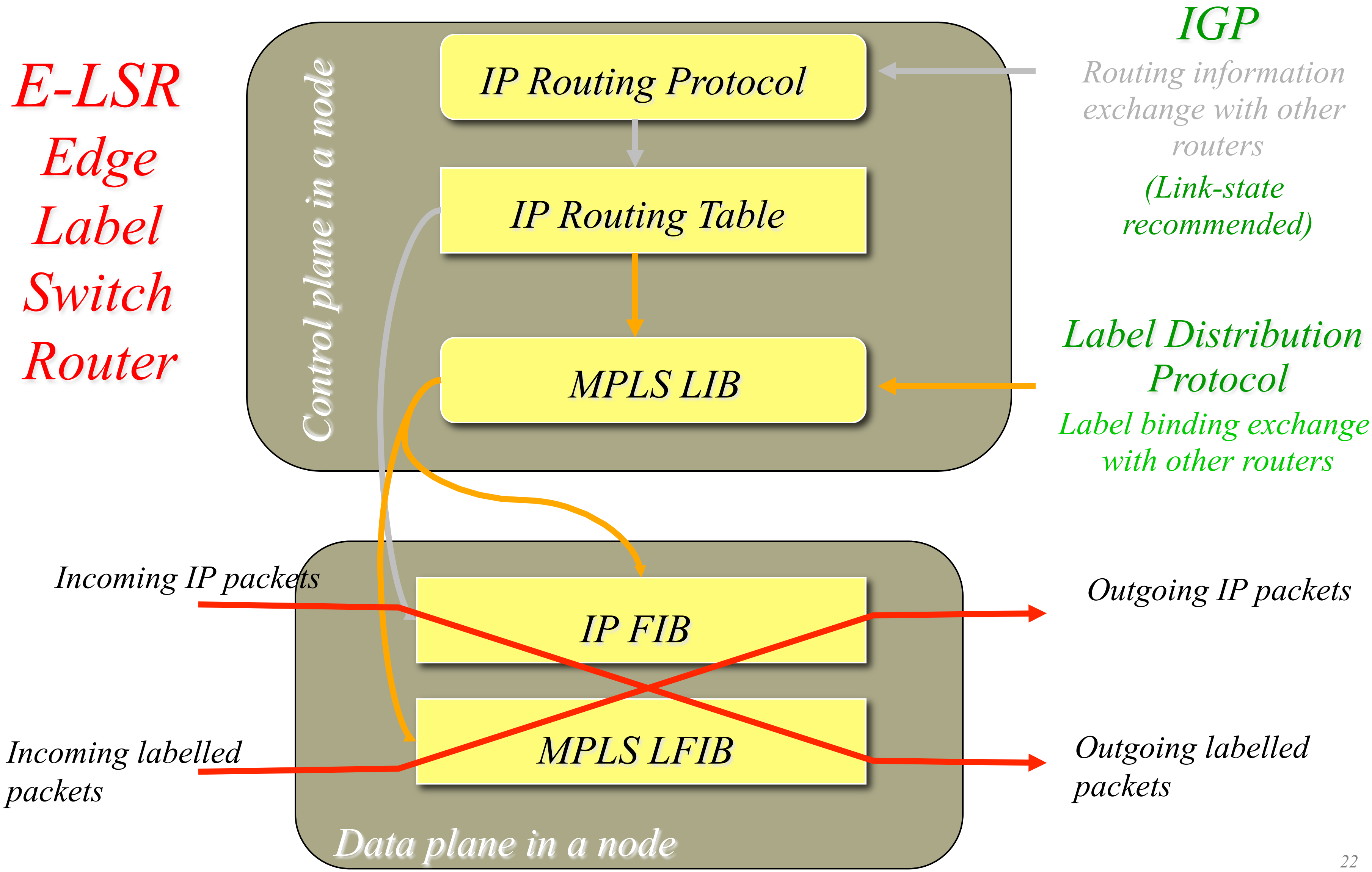
- also referred to as the control plane
- responsible for creating and maintaining label forwarding information (known as label bindings)
- label mappings distributed via some signaling protocol, e.g.

Label Distribution Protocol (LDP) or via extensions of BGP and RSVP (i.e. MP-iBGP and RSVP-TE resp.)

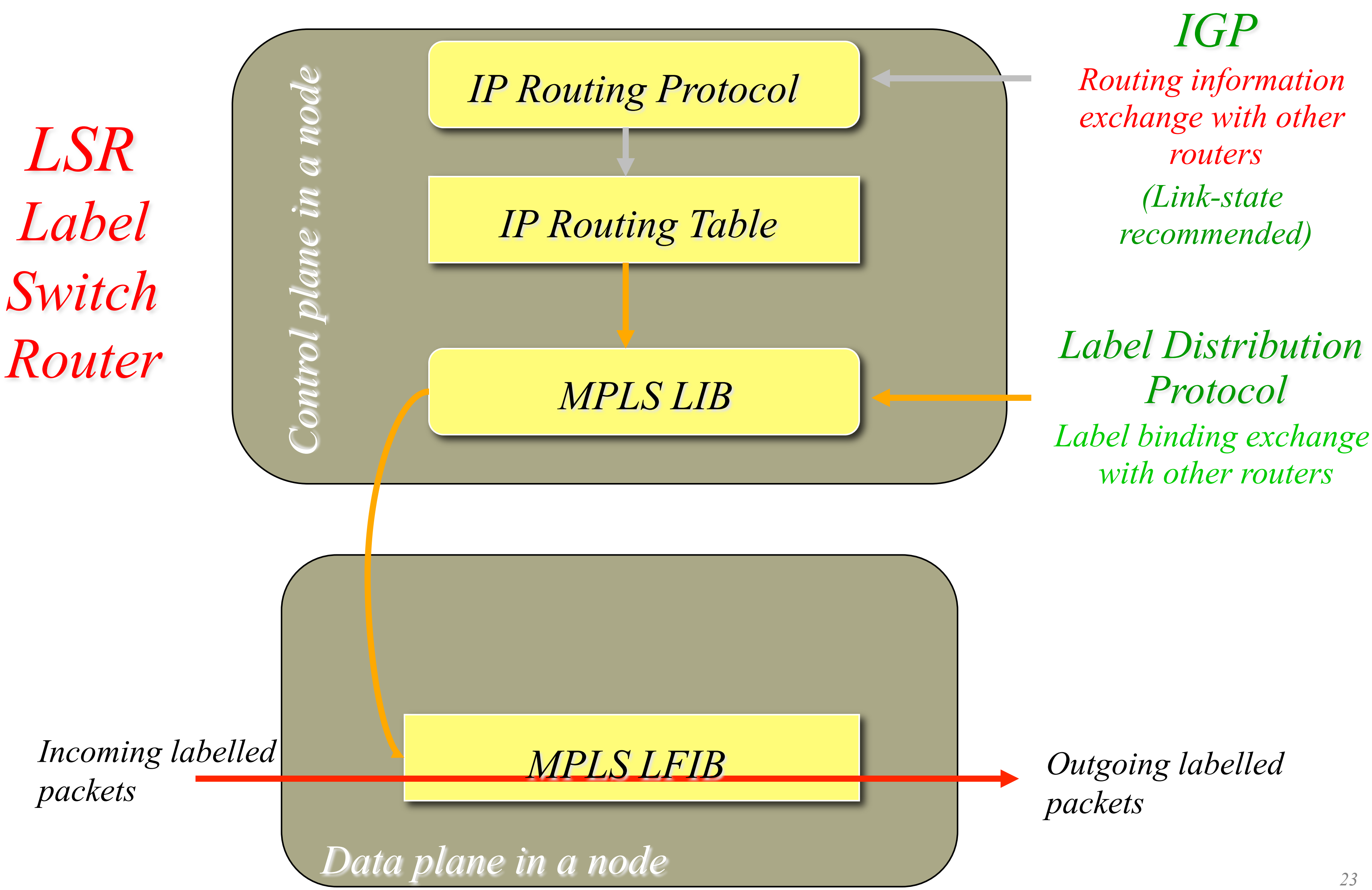
ISIS and OSPF also got extended to carry supp. info to support QoS-based, non-shortest-path routing in MPLS

The Intelligence

Control-Plane to Data-Plane
MPLS / E-LSR



Control-Plane to Data-Plane Core (i.e. non-edge) MPLS / LSR



MPLS Specific Tables

- Each LSR will use a LIB

  Label Information Base

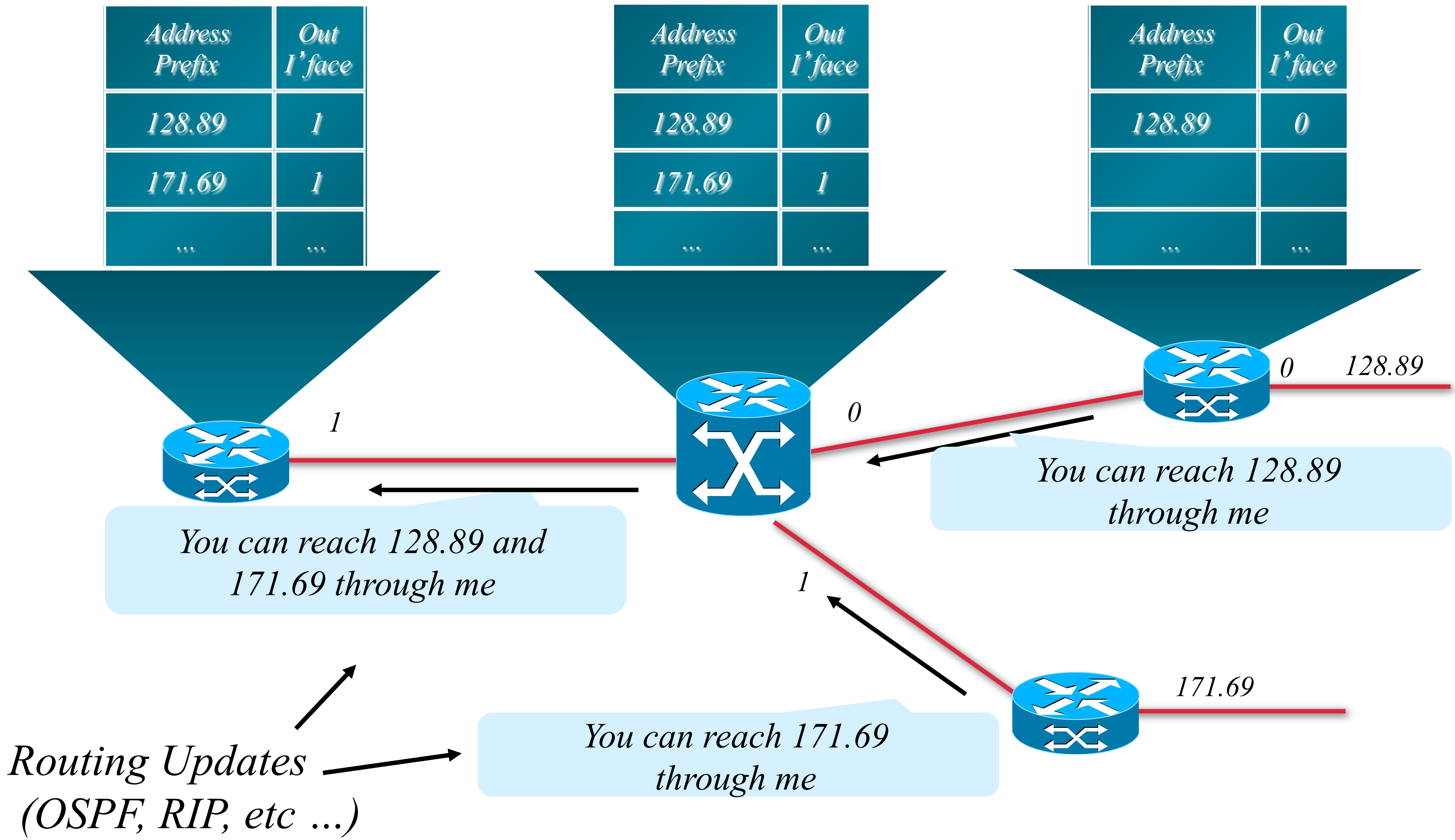
  Contains all label/prefix mappings from all LDP neighbours

- Each LSR will also use a LFIB

  Label Forwarding Information Base

  Contains only label/prefix mappings that are currently in use for label forwarding

MPLS Example: Routing Information

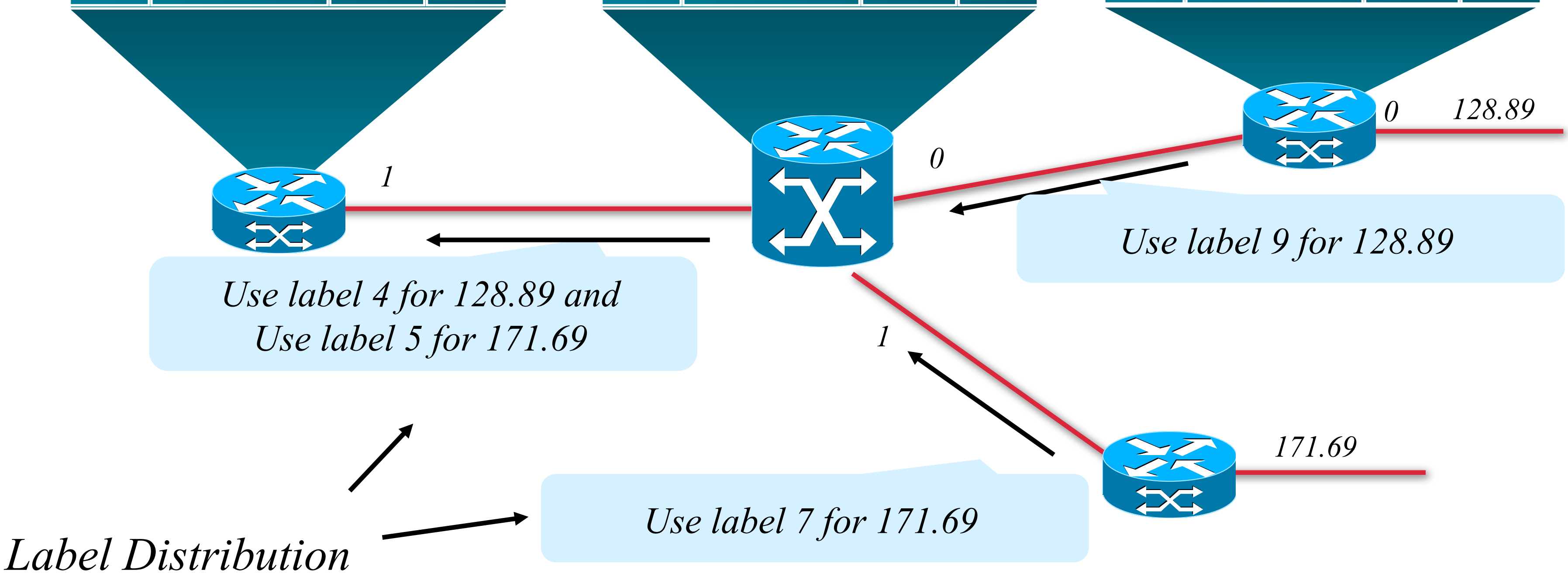


MPLS Example: Assigning Labels

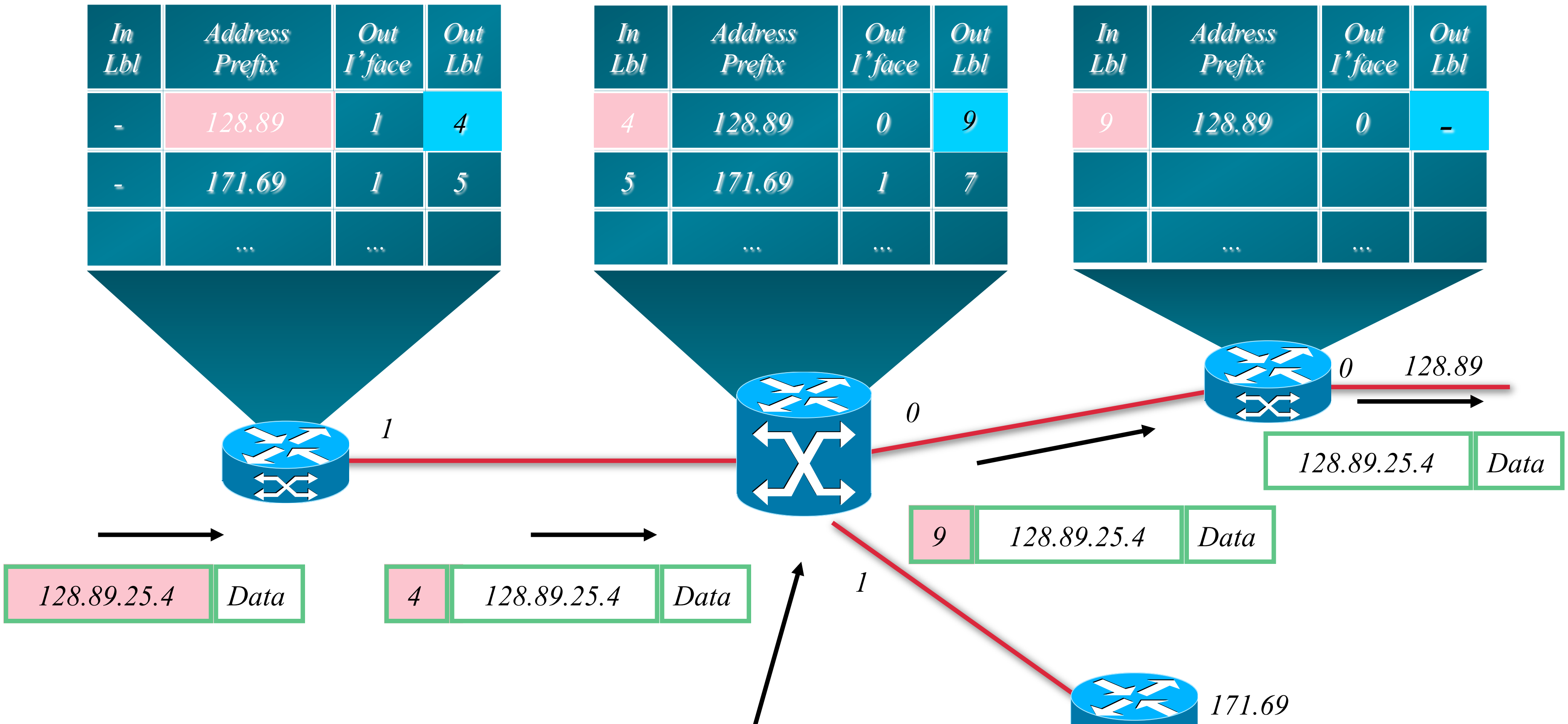
In Lbl	Address Prefix	Out I'face	Out Lbl
-	128.89	1	4
-	171.69	1	5
	...	...	

In Lbl	Address Prefix	Out I'face	Out Lbl
4	128.89	0	9
5	171.69	1	7
	...	...	

In Lbl	Address Prefix	Out I'face	Out Lbl
9	128.89	0	-
	...	...	

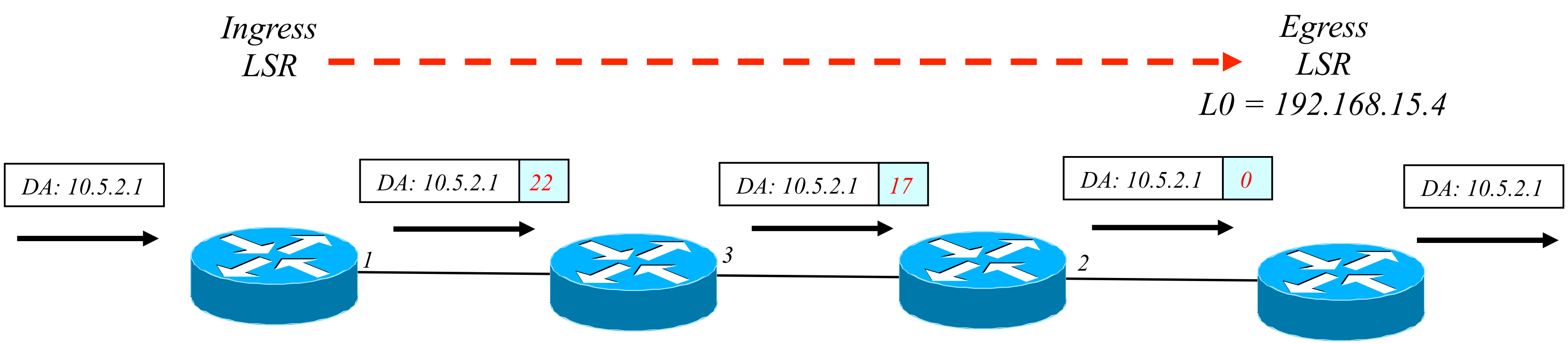


MPLS Example: Forwarding Packets



Label Switch Forwards Based on Label

Forwarding Mechanism seen as Label Pushing, Swapping, Popping



Routing Table

Prefix	Next Hop
10.5.0.0/16	192.168.15.4 PUSH 22, IF 1

MPLS Switching Table

IN	OUT
22	SWAP 17, IF 3

MPLS Switching Table

IN	OUT
17	SWAP 0, IF 2

MPLS Switching Table

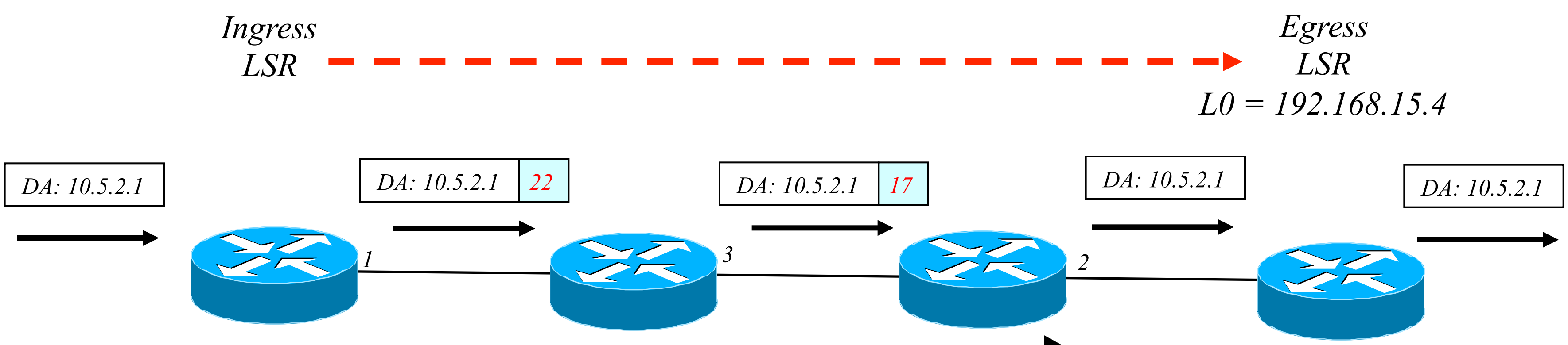
IN	OUT
0	POP

Routing Table

Prefix	Next Hop
10.5.0.0/16	10.1.16.3

- Label 0 = Explicit Null

Penultimate Hop Popping



Routing Table

Prefix	Next Hop
10.5.0.0/16	192.168.15.4 PUSH 22, IF 1

MPLS Switching Table

IN	OUT
22	SWAP 17, IF 3

MPLS Switching Table

IN	OUT
17	3, POP, IF 2

Routing Table

Prefix	Next Hop
10.5.0.0/16	10.1.16.3

- Label 3 = Implicit Null

Penultimate LSR:
Last transit LSR before egress

Label Encapsulation

Packet-over-SONET/SDH

Ethernet: similar

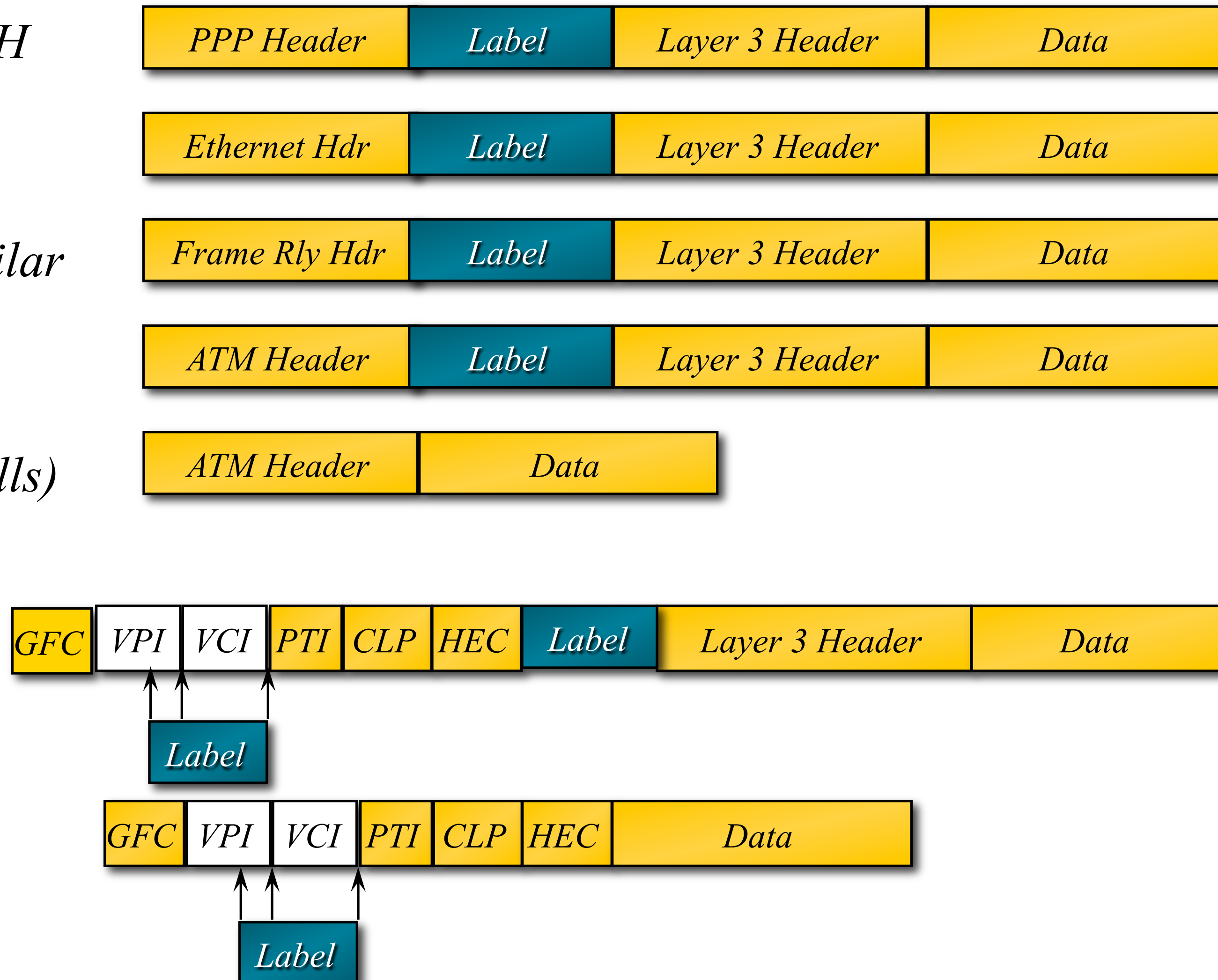
Frame Relay PVCs: similar

Label over ATM PVCs

(subsequent cells)

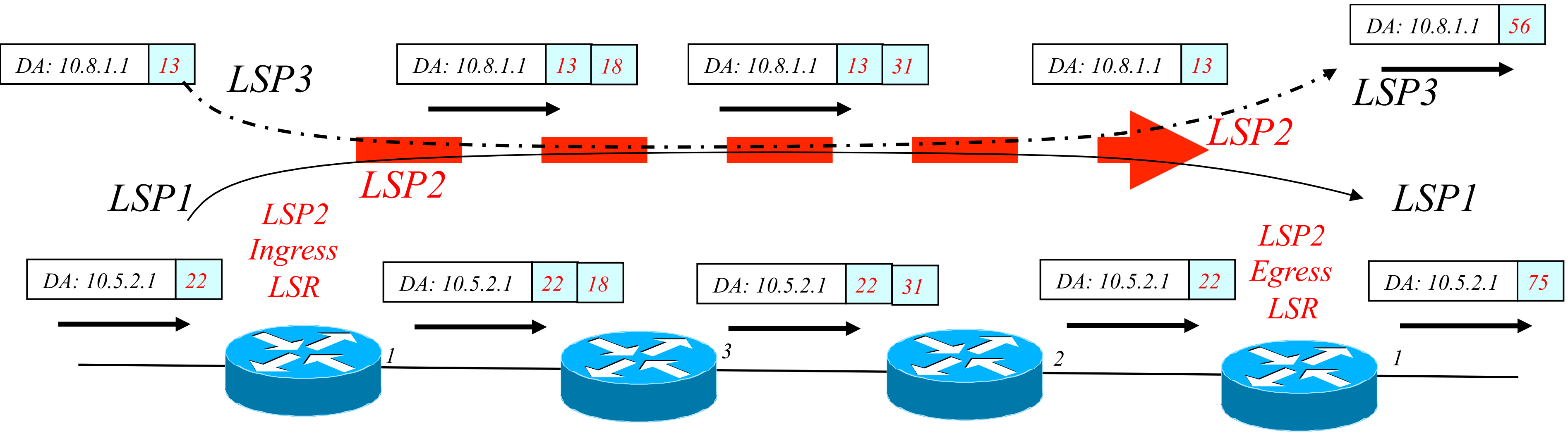
ATM label switching

(subsequent cells)



Label Stacking

- *Label Stacking allows LSPs to be tunneled (recursively) in other LSPs*
- *Labeled packet is forwarded based on the label at the top of the stack*



MPLS Switching Table

IN	OUT
22	PUSH 18, IF 1

MPLS Switching Table

IN	OUT
18	SWAP 31, IF 3

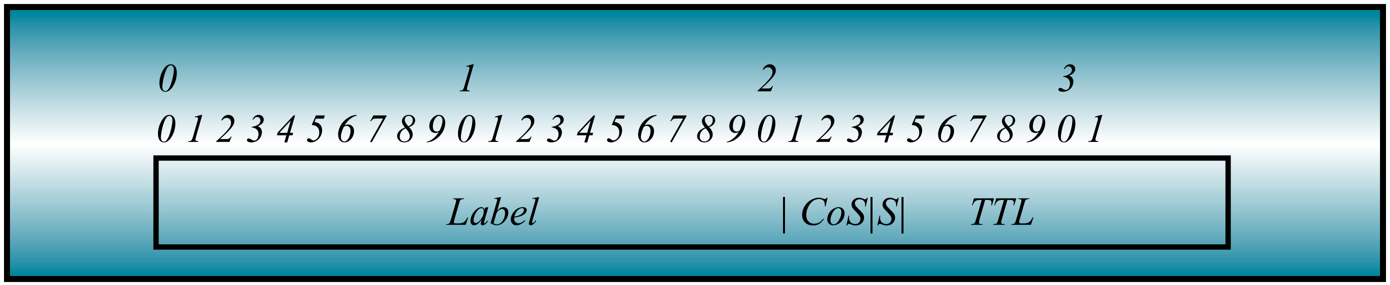
MPLS Switching Table

IN	OUT
31	POP , IF 2

MPLS Switching Table

IN	OUT
22	SWAP 75, IF 1

Label Encapsulation



- Label header is equal to 4 octets
  - Label value is 20 bits
  - Experimental is 3 bits
  - S (bottom of stack) is 1 bit
  - TTL (Time to live) is 8 bits

Label Values

0 - 15 Reserved

LABEL	DESIGNATION
0	IPv4 Explicit Null
1	Router Alert
2	IPv6 Explicit Null
3	Implicit Null
4-14	Reserved for Future Use
15	OAM
16 - 2^{20-1}	Production Use

What are the possible ways to allocate Labels ?

Downstream label allocation

label allocation is done by the downstream LSR

most natural mechanism for unicast traffic

Upstream label allocation

label allocation is done by the upstream LSR

may be used for optimality for some multicast traffic

A unique label for an egress LSR within the MPLS domain

Any stream to a particular MPLS egress node could use the label of that node.

Label Distribution

- Requests for labels flow downstream

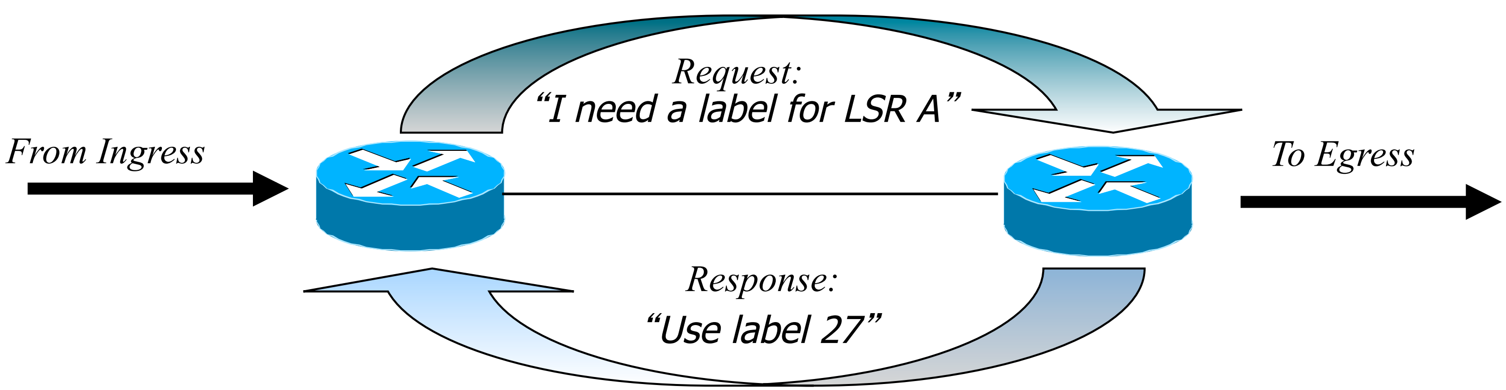
  Ingress ==> Egress

  Because ingress is the LSR that established the LSP

- Assignment of labels (label binding) flows upstream

  Egress ==> Ingress

  Because LSRs need to map *incoming* labels to some action (Push, Swap, Pop)



Most Common Label Distribution Modes in practice

- Downstream-on-Demand

  LSR requests its next hop for a label for a particular FEC

- Downstream Unsolicited

  LSR distributes bindings to LSRs that have not explicitly requested them

  For example, topology driven

  Only LDP and MPLS-BGP support Downstream Unsolicited mode

Possible ways to distribute/withdraw Labels

- **Use an explicit protocol, e.g. LDP**

  Separate routing computation and label distribution.

- **Piggybacking on Other Control Messages**

  Use existing routing/control protocol for distributing routing/control and label information, e.g. BGP, RSVP

- **Label purge mechanisms**

  By time out

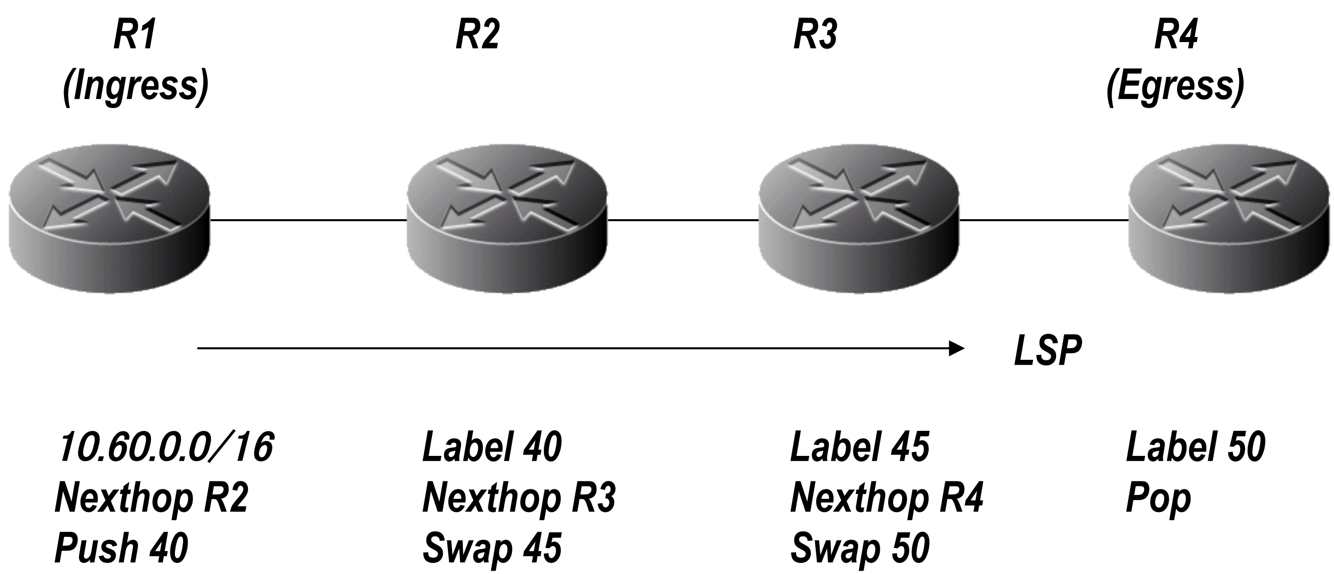
  Exchange of MPLS control packets

Label Distribution methods in practice

- **There are a number of possible label distribution methods:**
  - **Manual**
  - **MPLS-BGP (MP-iBGP-4) RFC2547, RFC4364**
  - **Resource Reservation Protocol-Traffic Engineering (RSVP-TE) (RFC 2205, RFC 2210)**
  - **Label Distribution Protocol (LDP) RFC3036, 5036**
  - **Constraint-Based LDP (CR-LDP) RFC3212, 3468<- lost battle with RSVP-TE, not used widely**

Manual Configuration

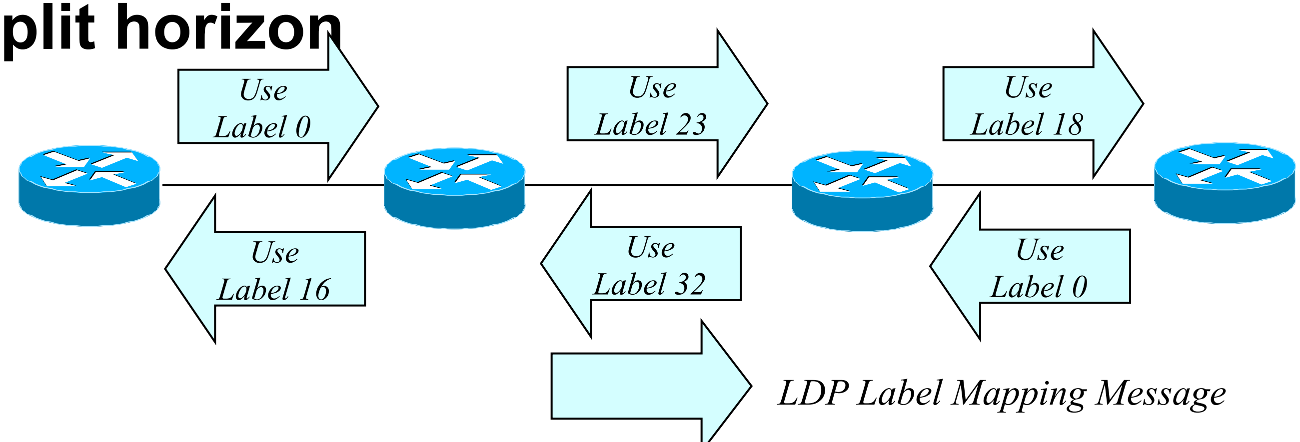
- **Labels are manually configured**
- **Useful in testing or to get around signaling problems**



The Label Distribution Protocol (LDP) RFC3036,5036

- Hop-by-hop label distribution
- Always follows IGP best path
- IP addresses are locally bound to labels
- Bindings are stored in Label Information Base (LIB)
- All bindings advertised to all peers

No split horizon



LDP (cont'd)

- **Supports Downstream on Demand and Downstream Unsolicited**
- **No support for QoS or traffic engineering**
- **UDP used for peer discovery**
- **TCP used for session, advertisement and notification messages**
- **Uses Type-Length-Value (TLV) encoding**
- **Highly scalable**
    - **Best suited for apps using thousands of LSPs (VPNs)**

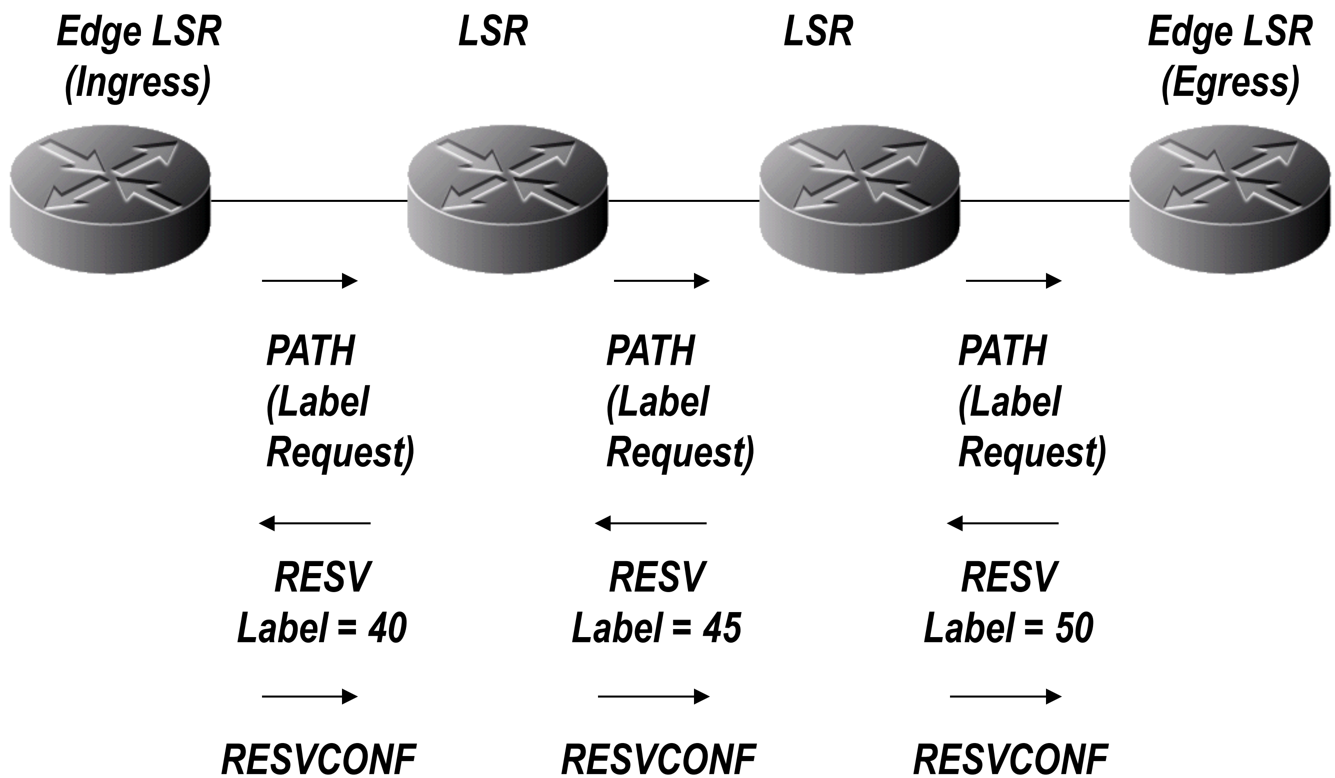
MPLS-BGP

- Use MP-iBGP-4 to distribute label information as well as VPN routes
- BGP peers can send route updates and the associated labels at the same time
- Route reflectors can also be used to distribute labels to increase scalability

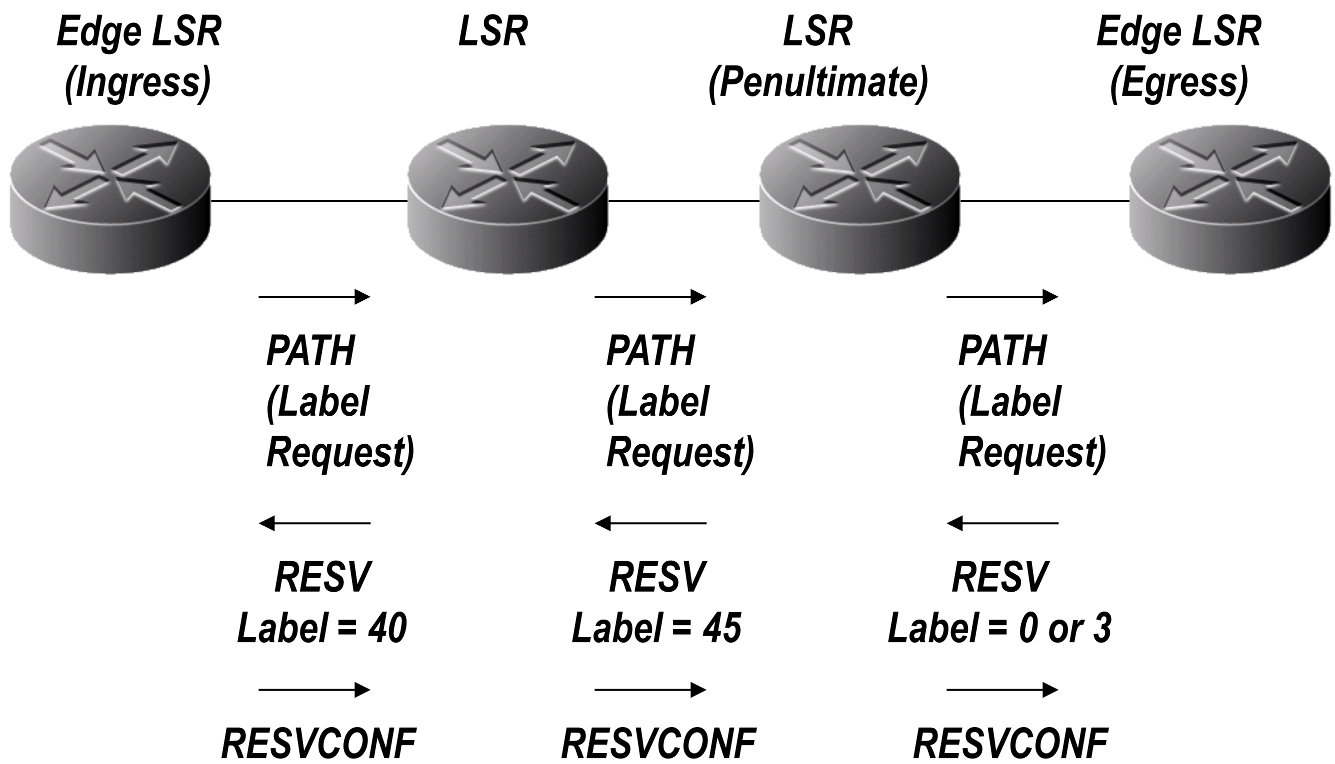
RSVP-TE

- Traffic Engineering (TE) extensions added to RSVP (Resource Reservation Protocol)
  - Sender and receiver are ingress and egress LSRs
  - New objects have been defined
- Supports Downstream on Demand label distribution
- PATH messages used by sender to solicit a label from downstream LSRs
- RESV messages used by downstream LSRs to pass label upstream towards the sender
- Less scalable -- LSRs maintain soft state ; need periodic refresh of PATH/RESV messages
  - Best suited for traffic engineering in the core

RSVP-TE Operation



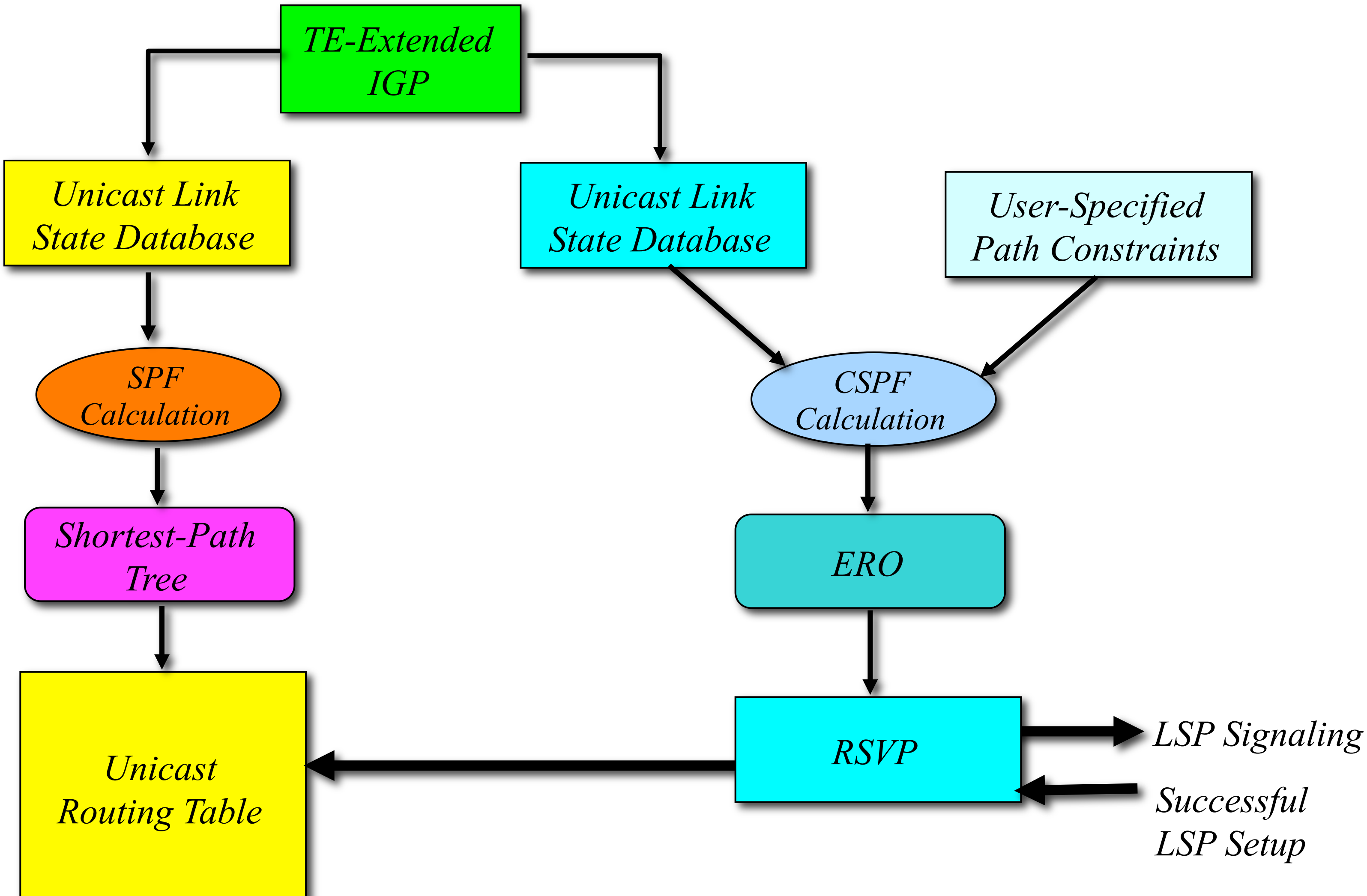
RSVP-TE Operation with PHP



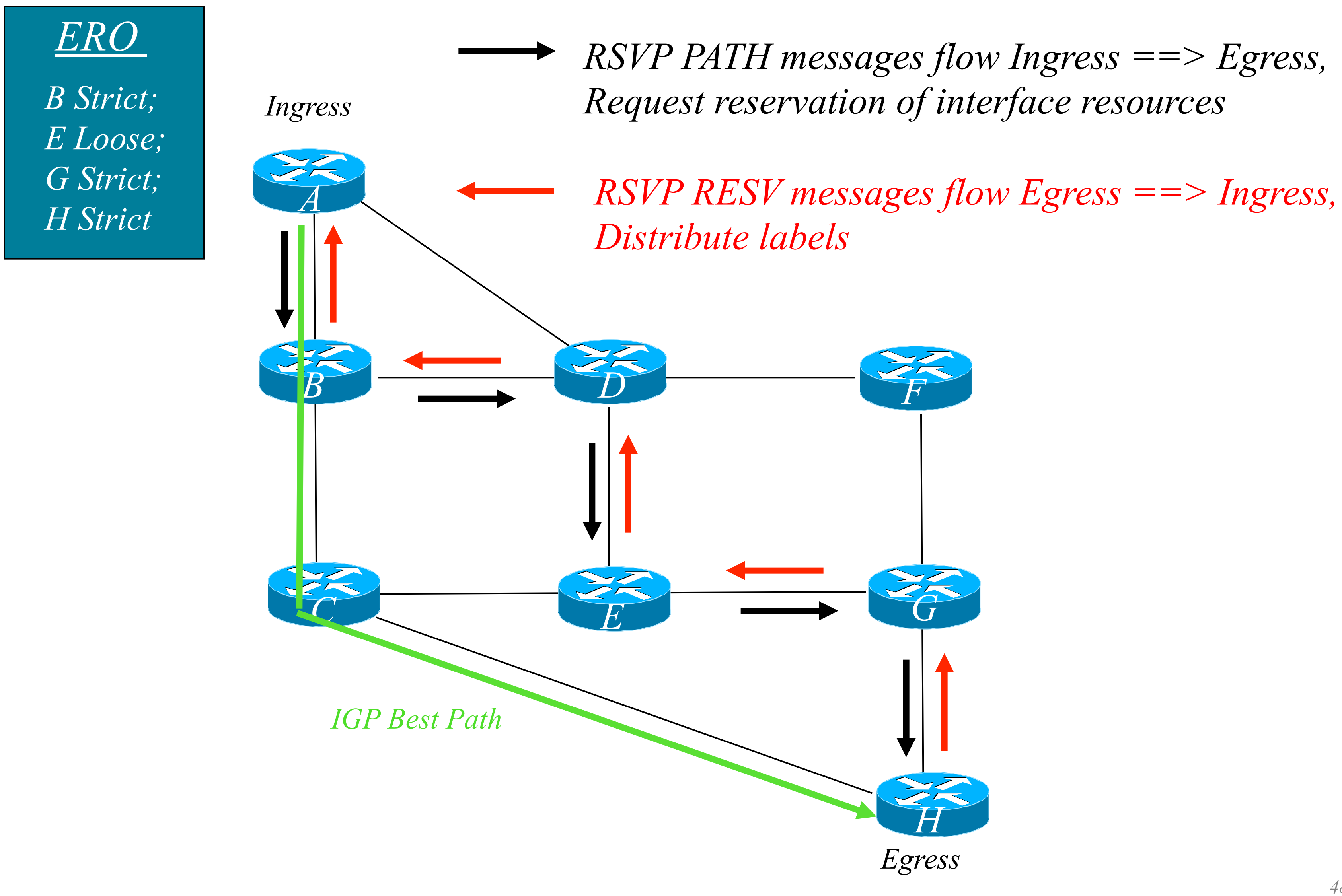
Label Distribution: RSVP-TE

- **Support End-to-end *constrained* path signaling**
- **Enabled by OSPF or IS-IS with TE extensions**
  - **Extended IGPs flood TE interface parameters:**
    - **Maximum Bandwidth**
    - **Maximum Reservable Bandwidth**
    - **Unreserved Bandwidth**
    - **TE Metric**
    - **Administrative Group (aka Link Affinity or "Link Coloring")**
  - **OSPF uses opaque LSA and IS-IS uses new TLV to carry TE-info**
- **Interface parameters used to build *Traffic Engineering Database* (TED)**
- ***Constrained Shortest Path First* (CSPF)**
  - **Calculates best path based on specified constraints**
- ***Explicit Route Object* (ERO) passed to RSVP**

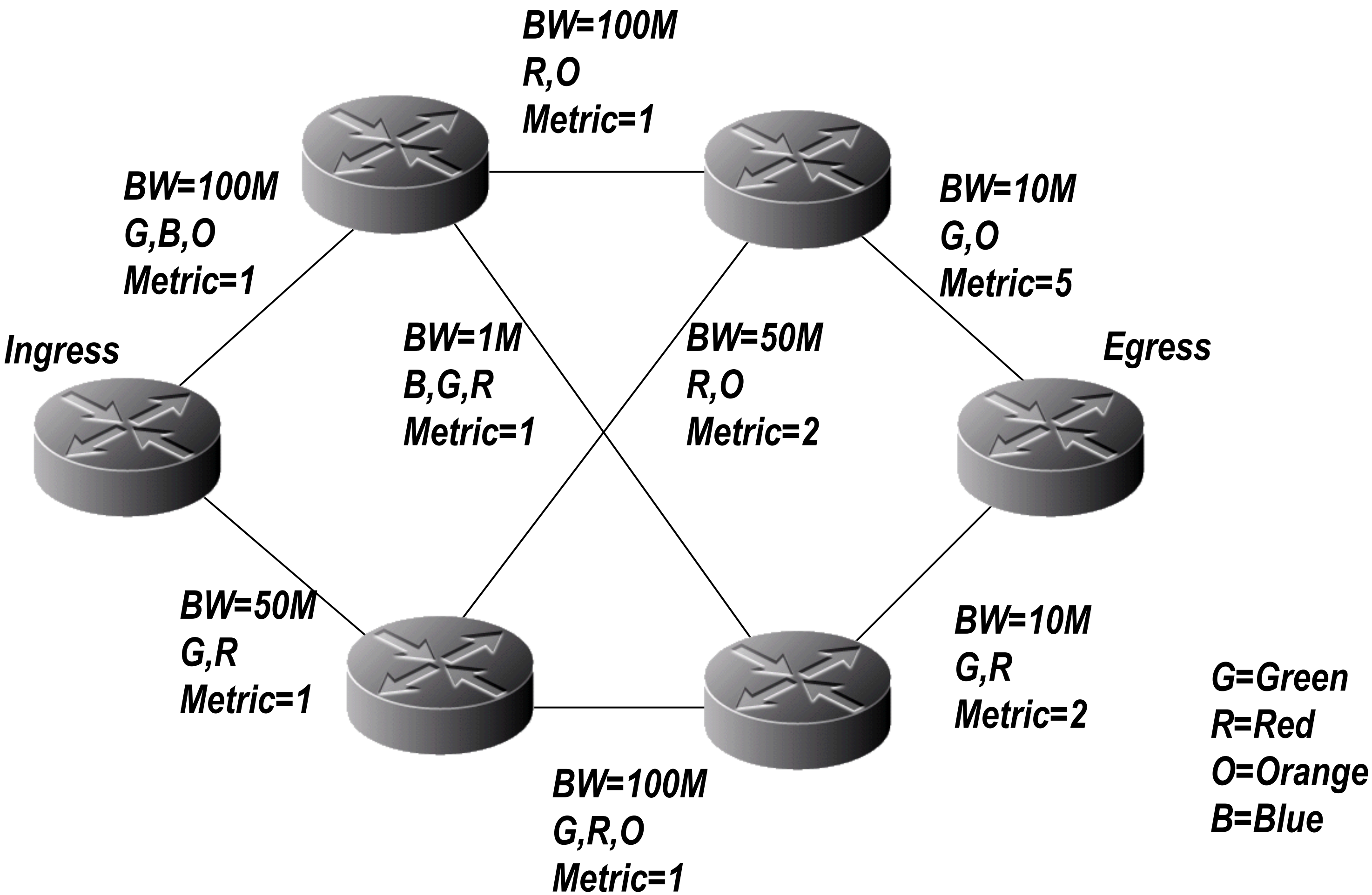
CSPF Calculation



RSVP-TE LSP Signaling

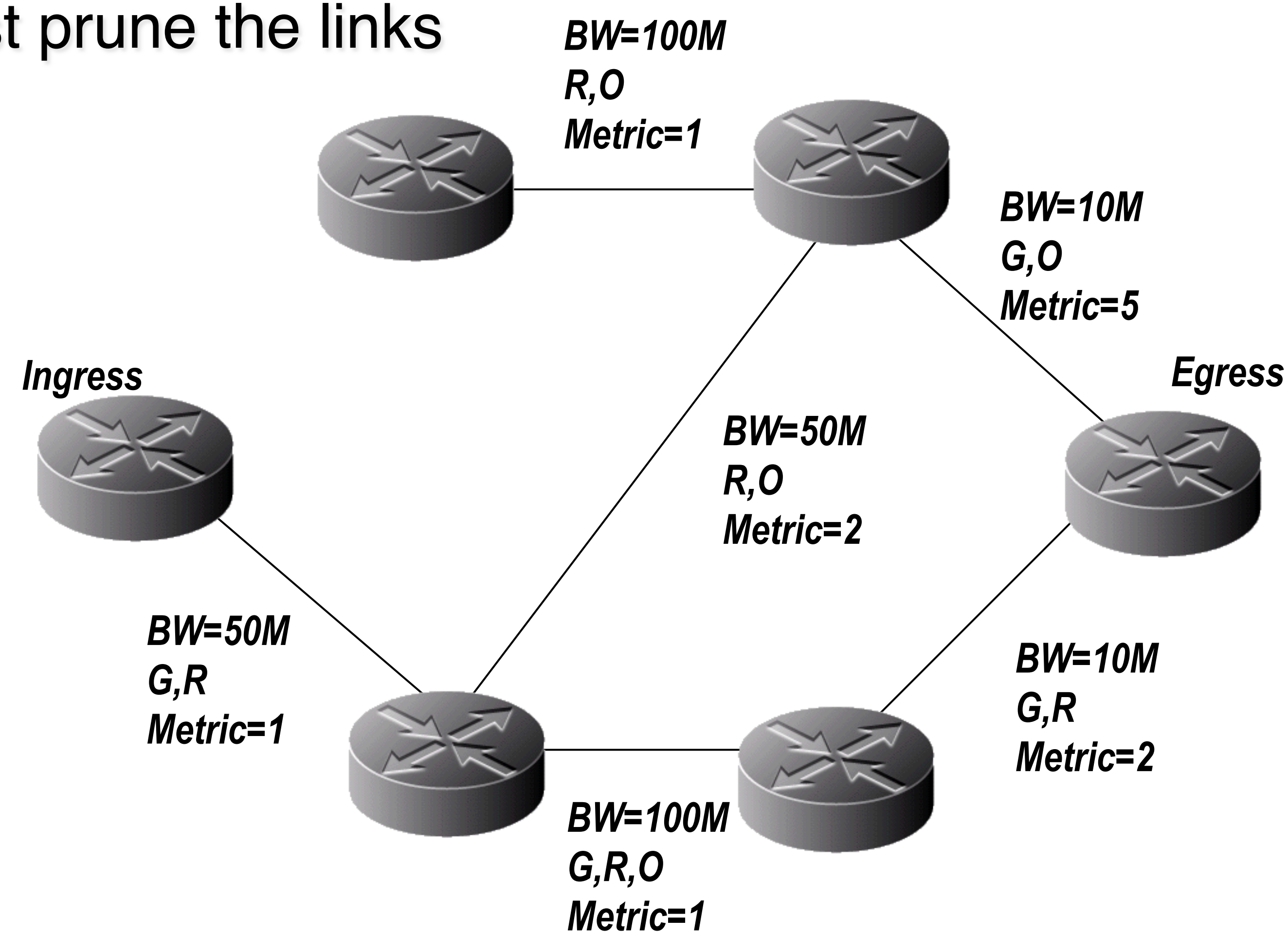


An Example: Traffic Engineering Database



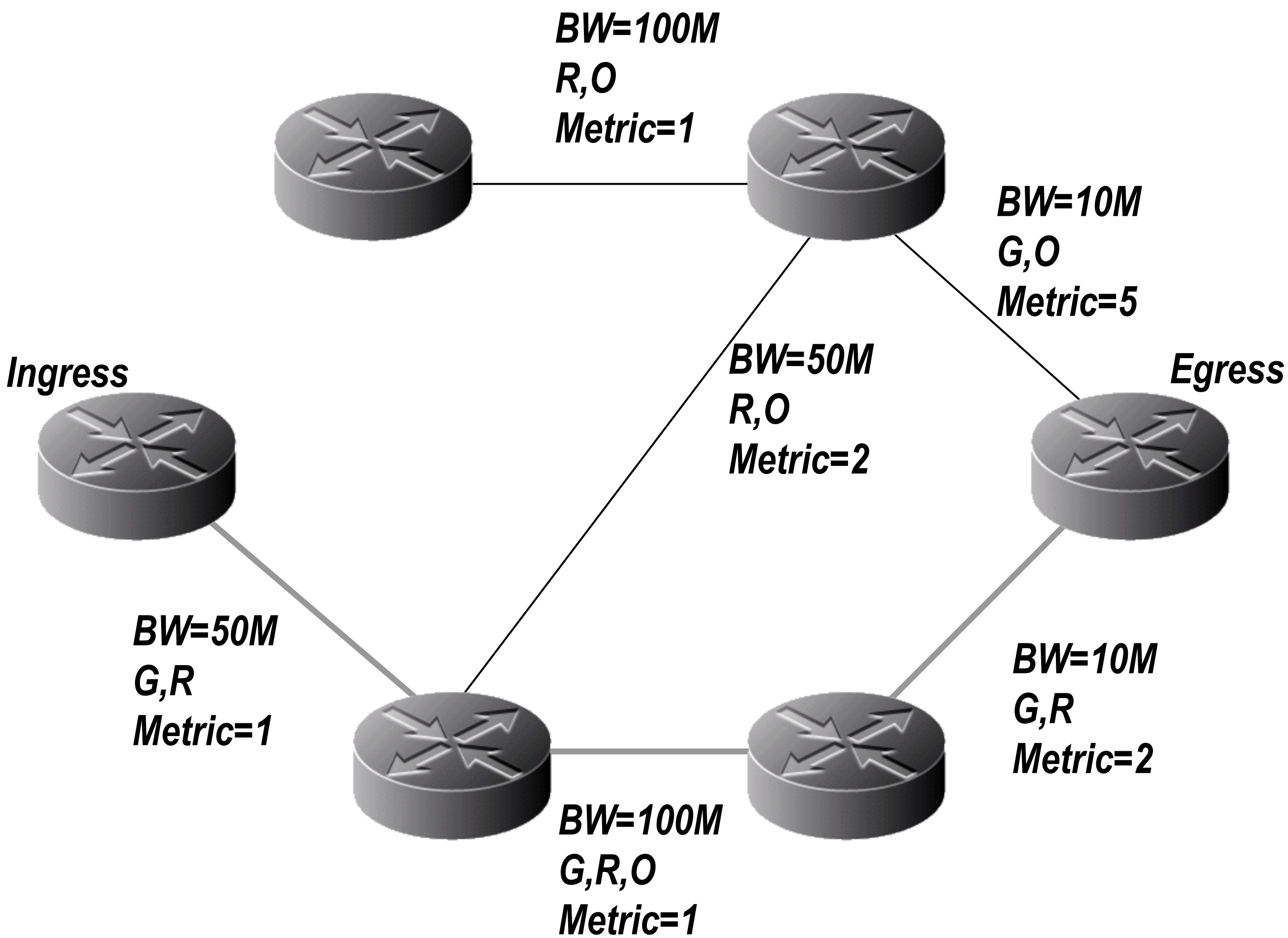
Selecting a Path

- How to select a 2M path which excludes any blue links?
- First prune the links



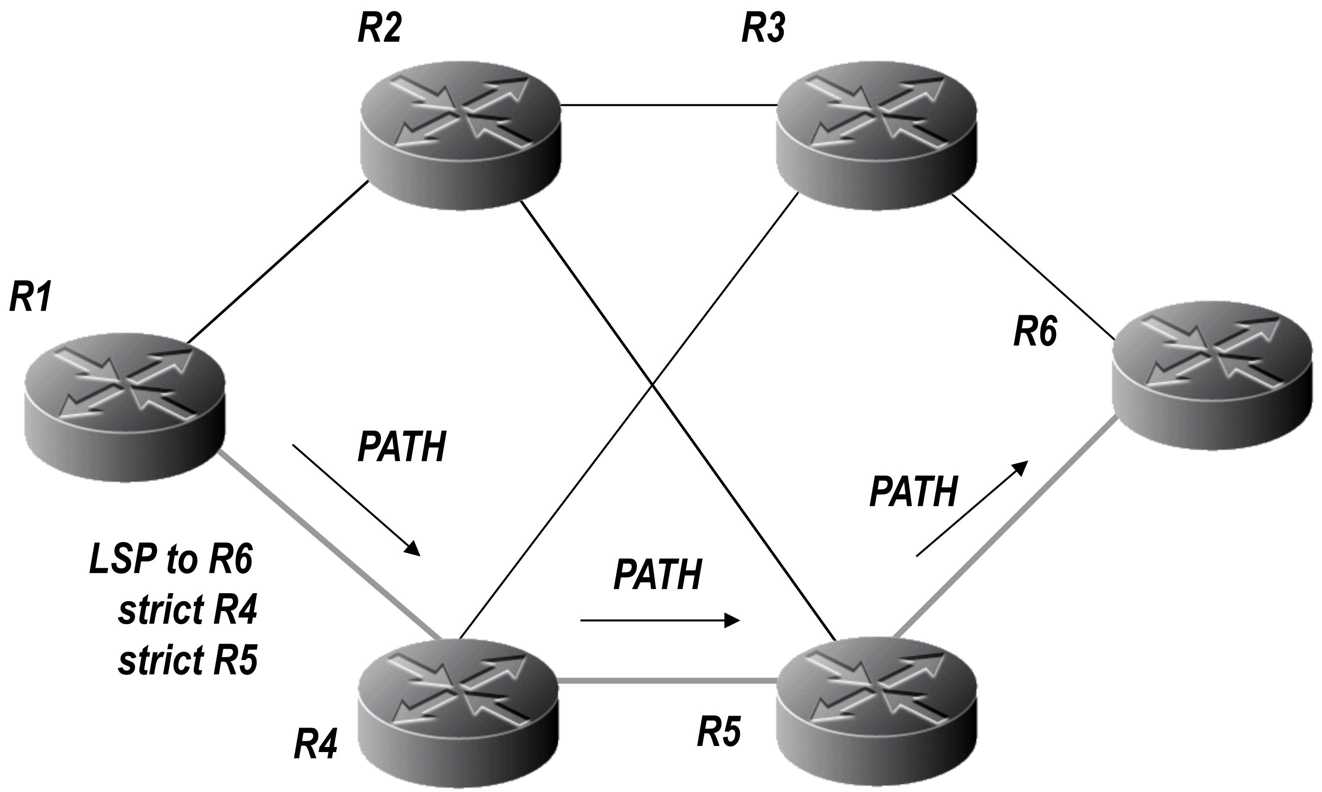
Selecting a Path

- Now select the shortest path



Explicit Route

- Once the path has been determined, the ingress router will typically signal the path using the Explicit Route Option (ERO) or ER-TLV



RSVP-TE and LDP Applications

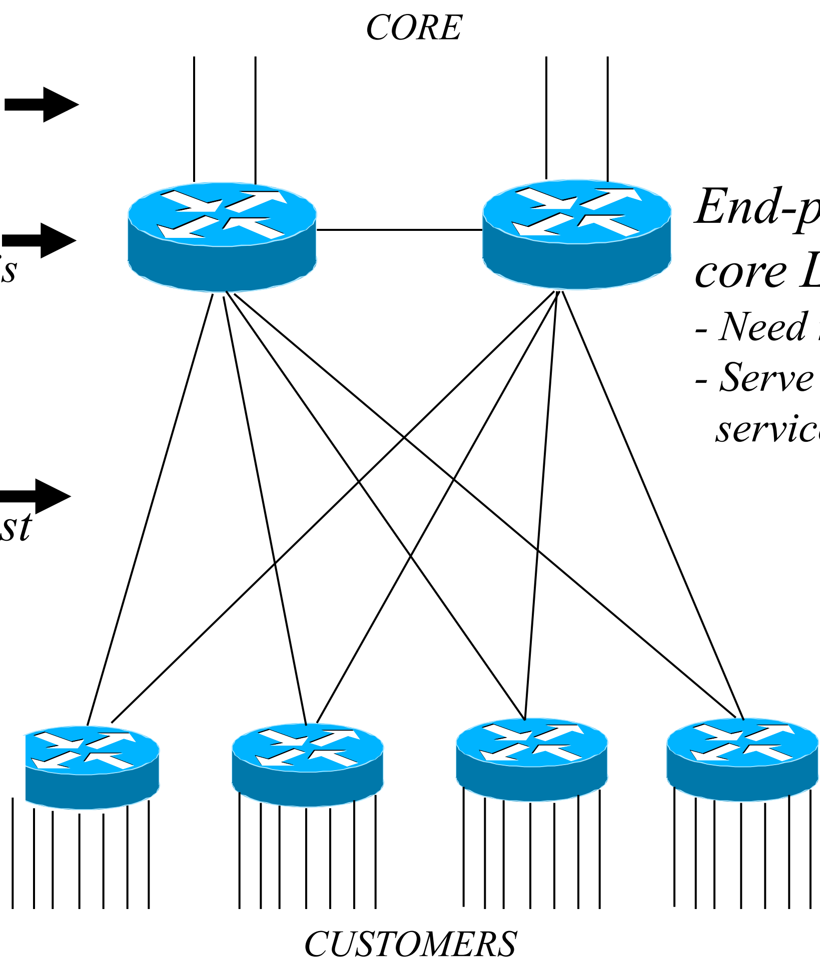
Typical PoP architecture:

High-bandwidth core uplinks

Core routers
Primary requirement is high performance

Aggregation routers might or might not exist here

Edge routers
Primary requirements are:
- *Service intelligence*
- *Customer-facing interface density*



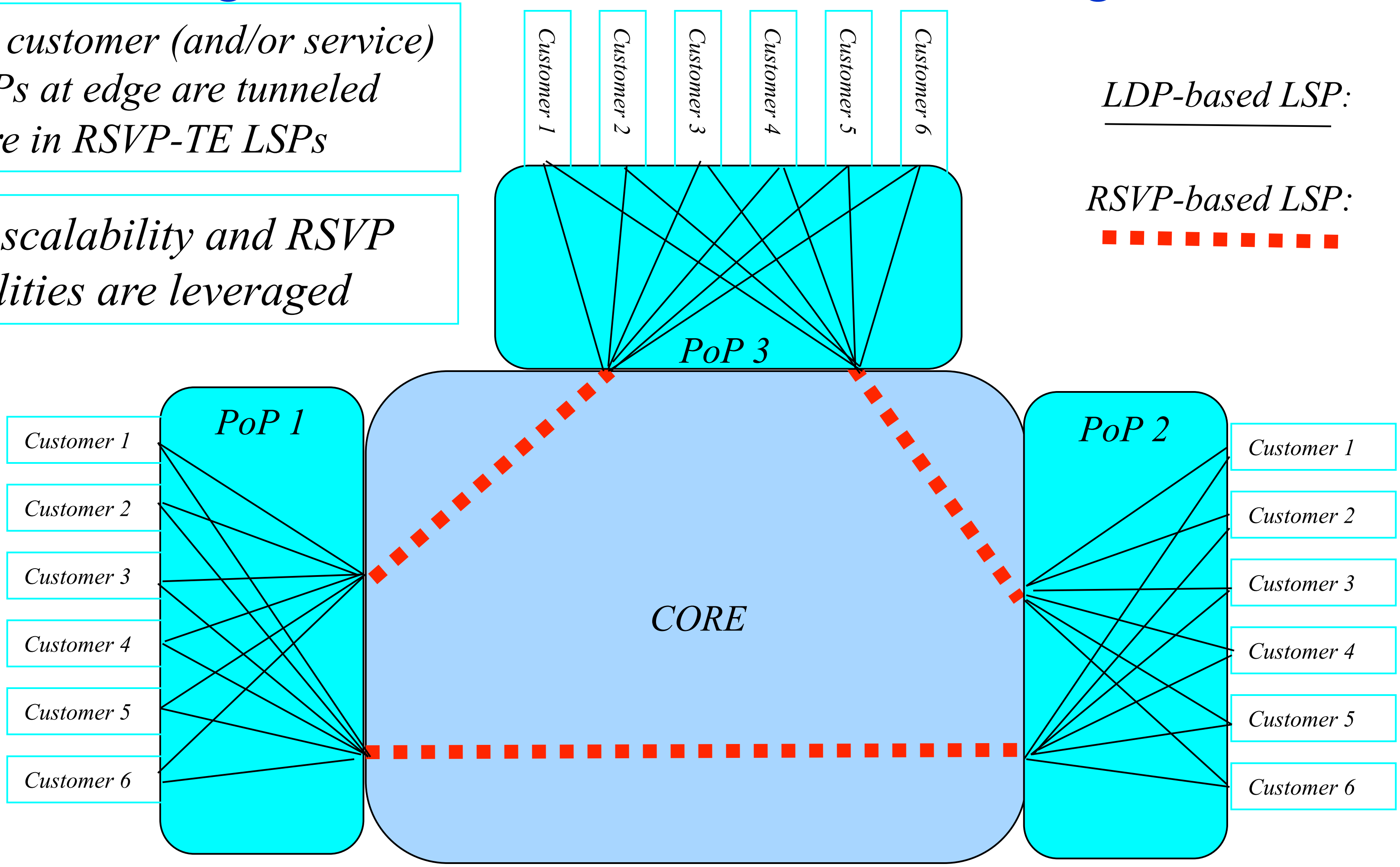
End-point for RSVP-TE core LSPs
- *Need relatively few of these*
- *Serve as PoP-to-PoP tunnels for service-specific edge LSPs*

End-point for LDP service-specific LSPs
Might be hundreds or thousands of these

Using RSVP-TE and LDP LSPs Together

LDP-based customer (and/or service) specific LSPs at edge are tunneled through core in RSVP-TE LSPs

Both LDP scalability and RSVP TE capabilities are leveraged



Summary: Benefits of MPLS

- Benefits relative to use of a Router Core
  - Simplified forwarding (avoid longest prefix match)
  - Efficient explicit routing
  - Traffic Engineering
  - QoS routing
  - Complex mappings from IP packet to forwarding equivalence class (FEC)
  - Partitioning of functionality: Control vs. Data Plane
  - Single forwarding paradigm with several level differentiation
- Benefits relative to use of an ATM or Frame Relay Core
  - Scaling of the routing protocol
  - Common operation over packet and cell media
  - Easier Management
  - Elimination of the ‘routing over Large Clouds’ issue

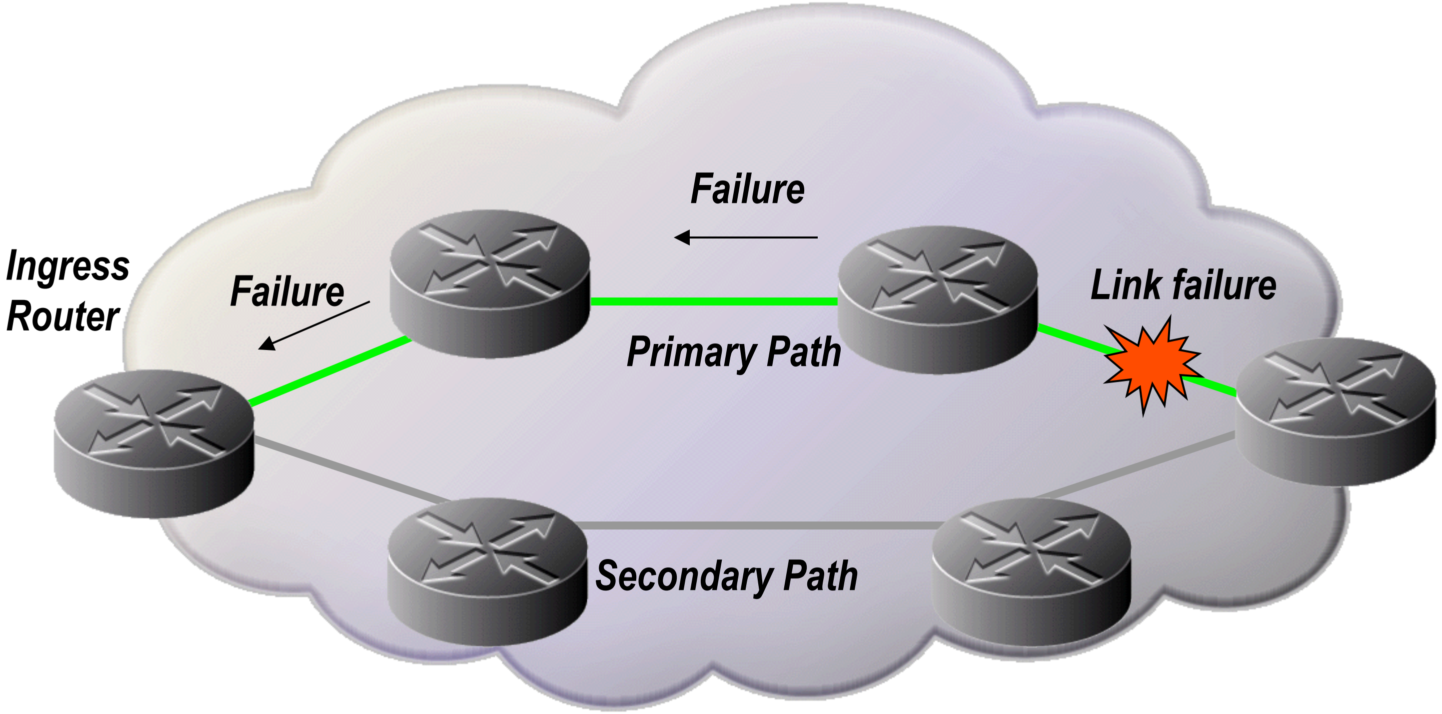
Sample Applications of MPLS

- Traffic engineering
  - QoS-based Routing along non-shortest paths
  - Can support FEC-specific forwarding (Differentiated services for different FECs)
- Enhanced Route Protection against Link and node failures
  - Fast restoration to an alternative LSP
- Virtual Private Networks (VPNs)
  - Layer 3 VPNs
  - Layer 2 VPNs, e.g. Virtual Private LAN Service (VPLS)
- This Idea subsequently generalized to support signaling-based "virtual-circuit" setup and TE in Optical Transmission Networks under the names: Multiple Protocol Lambda Switching, Generalized MPLS (GMPLS), and MPLS-Transport Profile (MPLS-TP)

Application Example: Enhanced Route Protection

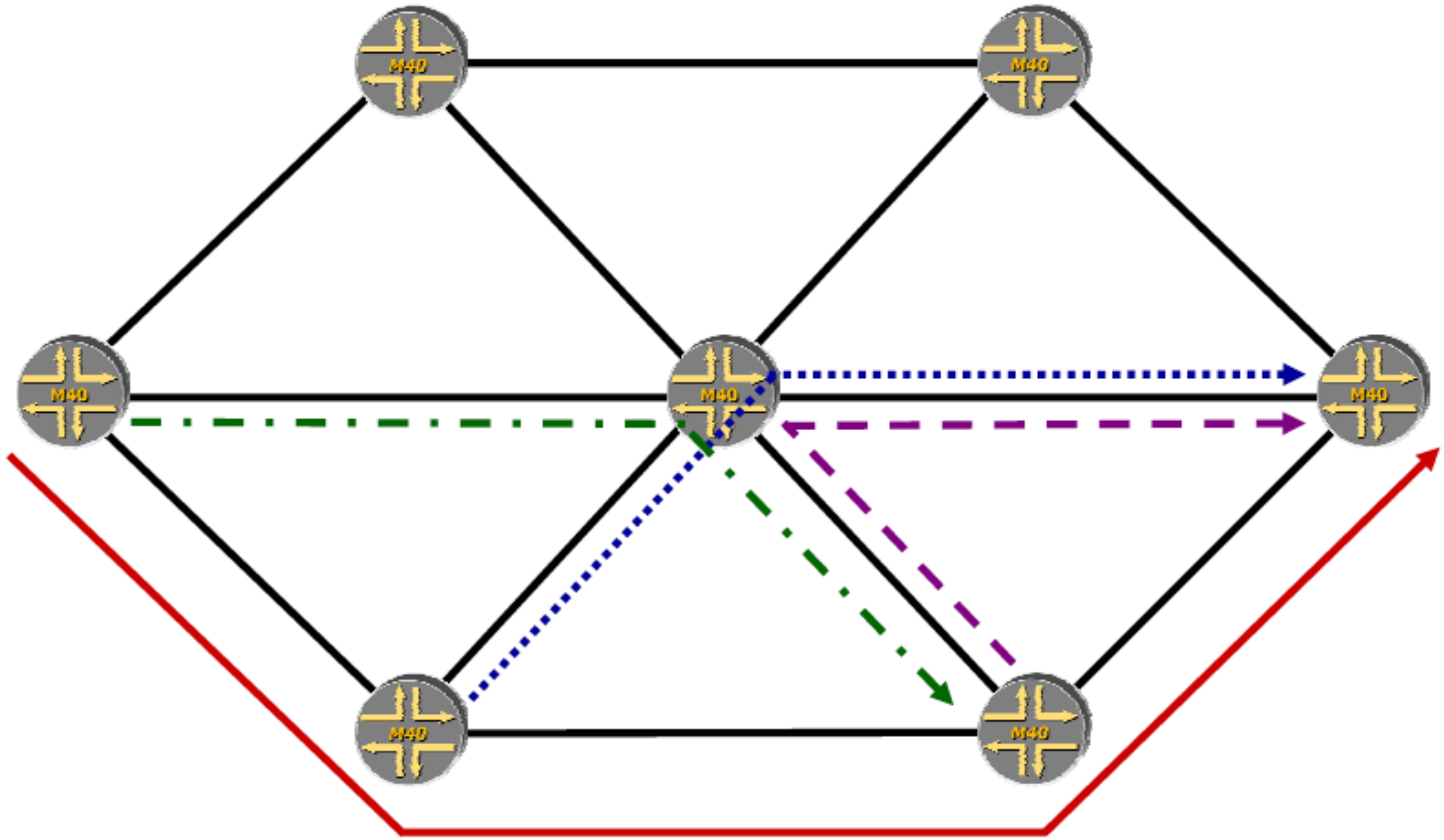
- Head-end Reroute
  - If a link along the path fails, the ingress node is notified
  - The ingress node must recompute another path and then set up the new path
- End-to-end Path-based Protection Switching
  - Pre-establish two paths for an LSP for redundancy
  - If a link along the primary path fails, the ingress node switches over to the secondary path
- Localized Fast Reroute for link & node protection
  - Each node pre-computes and pre-establishes a path to bypass potential failures in the downstream link or node

Example:
E2E Path-based Protection Switching



When ingress router is notified of the link failure, it switches all traffic to the secondary path.

Example: Localized Fast Reroute for Node & Link Protection



- Each node creates an alternate LSP around its downstream node (and the interconnecting link)
- Penultimate node uses link protection

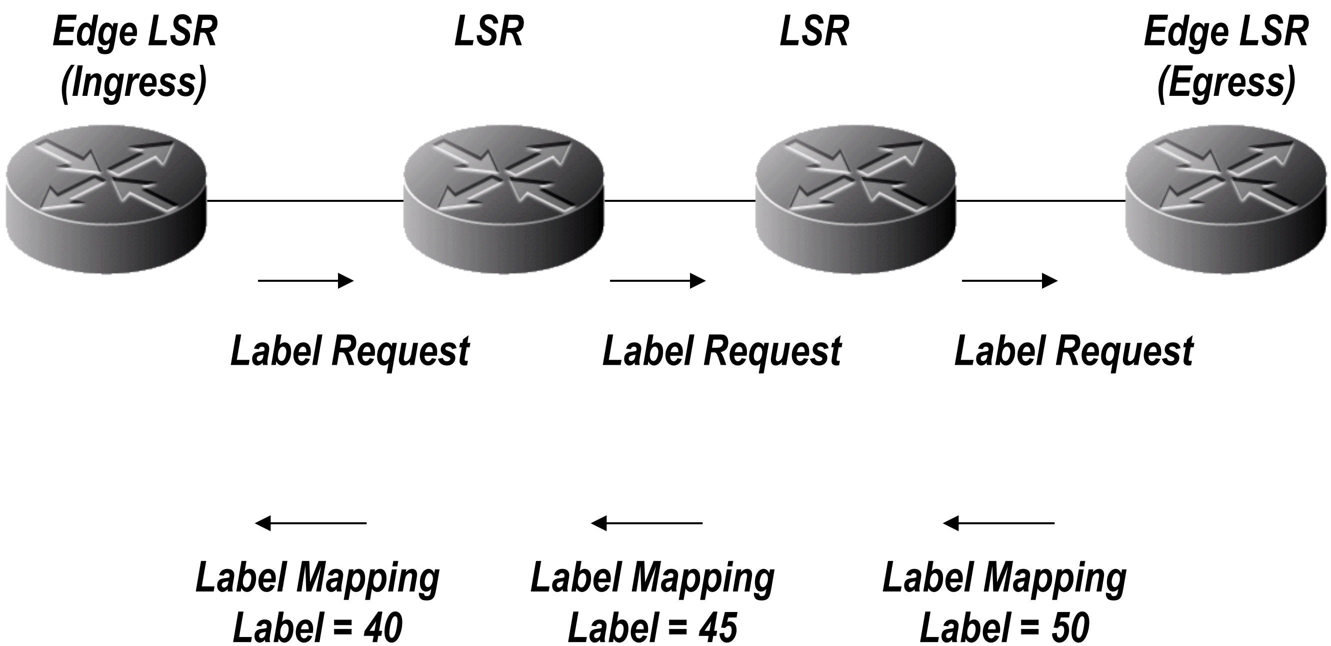
Backup Slides
on
CR-LDP (Deprecated)

ConstRaint-based LDP
(CR – LDP)

- Extensions to LDP that convey resource reservation requests for user and network constraints
- CR-LDP uses TCP sessions between LSR peers to send LDP messages
- A mechanism for establishing explicitly routed LSPs
- An Explicit Route is a Constrained Route

  Ingress LSR calculates entire route based on Traffic Engineering Database (TED) and known constraints

CR-LDP Operation



CR-LDP vs RSVP-TE

- **Signaling Attributes**
- **LSP Attributes**
- **Traffic Engineering Attributes**
- **Reliability & Security Mechanisms**

Signaling Attributes

	CR-LDP	**RSVP-TE**
Underlying Protocol	*LDP*	*RSVP*
Transport Protocol	*TCP*	*Raw IP*
Protocol State	*Hard*	*Soft*
Multipoint-to-Point	*Yes*	*Yes*
Multicasting	*No*	*No*

LSP Attributes

	CR-LDP	***RSVP-TE***
Explicit Routing	*Strict & Loose*	*Strict & Loose*
Route Pinning	*Yes*	*Yes*
LSP Re-Routing	*Yes*	*Yes*
LSP Preemption	*Yes*	*Yes*
LSP Protection	*Yes*	*Yes*
LSP Merging	*Yes*	*Yes*
LSP Stacking	*Yes*	*Yes*

Traffic Engineering Attributes

	CR-LDP	***RSVP-TE***
Traffic Control	*Forward Path*	*Reverse Path*

- **CR-LDP**
  - **Negotiates resources during the Request process**
  - **Confirms resources during the Mapping process**
  - **LSPs are setup only if resources are available**
  - **Ability exists to allow for negotiation of resources**

Traffic Engineering Attributes

	CR-LDP	***RSVP-TE***
Traffic Control	*Forward Path*	*Reverse Path*

- **RSVP-TE**
  - **Passes resource requirements to the Egress LER**
  - **Egress LER converts the Tspec into a Rspec**
  - **Resource reservations occur on RESV process**

Reliability & Security Attributes

	CR-LDP	***RSVP-TE***
Link Failure Detection	*Yes*	*Yes*
Failure Recovery	*Yes*	*Yes*
Security Support	*Yes*	*Yes*

RSVP-TE vs. CR-LDP

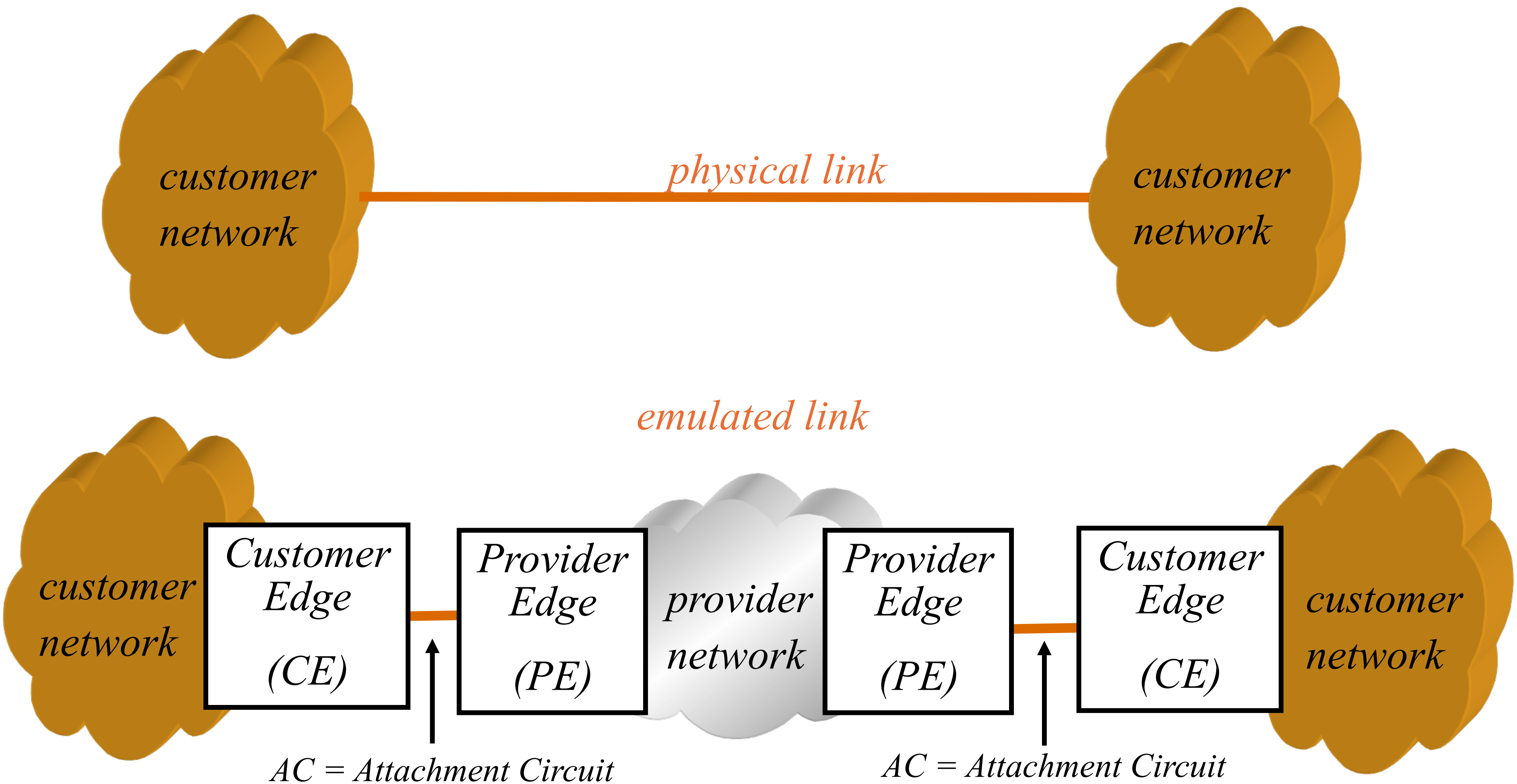
- **Each protocol has strengths & weaknesses**
- **CR-LDP is based upon LDP which supposed to give it an advantage of using a common protocol**
- **BUT**

  CR-LDP lost the battle, seldom deployed in practice ;

  RSVP-TE is used instead.

Layer 2 or Layer VPNs using MPLS

Basic L2 or L3 VPN model

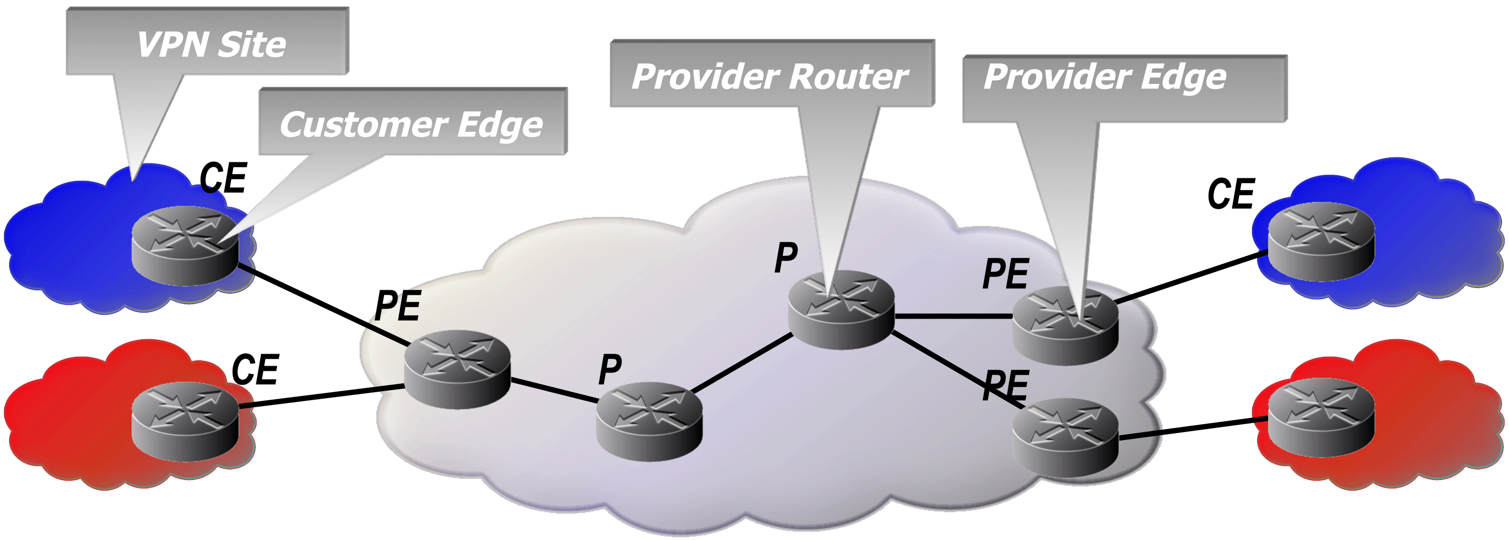


*provider network may be **L3** (e.g. IP) or **L2** (e.g. Ethernet) or MPLS*

MPLS-based L2 or L3 VPNs

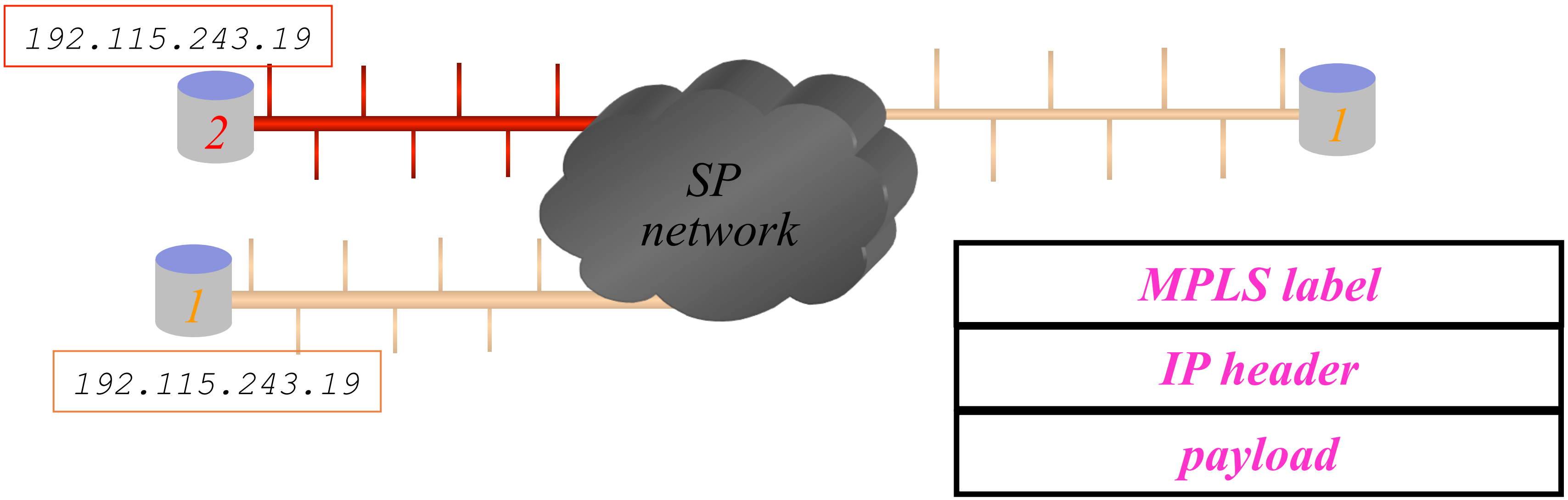
- MPLS can provide the required tunneling mechanism
  - MPLS can be used to provide traffic engineered PE-to-PE tunnels
  - An additional MPLS label can also be used to associated packets with a VPN
- VPNs based on delivering Layer 3 (IP) packets over MPLS tunnels are Layer 3 VPNs
  - RFC 4364 defined BGP/MPLS VPNs
- VPNs based on delivering Layer 2 (Ethernet) frames over MPLS tunnels are Layer 2 VPNs
  - Pseudo Ethernet Wire Service (PEWS) or Virtual Private Wire Service (VPWS)
  - RFCs 4761,4762 defined Virtual Private LAN Service (VPLS)

MPLS VPN Terminologies



- Customer Edge (CE) device: device located on customer premises
- Provider Edge (PE) device: maintains VPN-related information, exchanges VPN information with other Provider Edge devices, encapsulates/decapsulates VPN traffic
- Provider (P) router: forwards traffic VPN-unaware

MPLS solves IP address problem

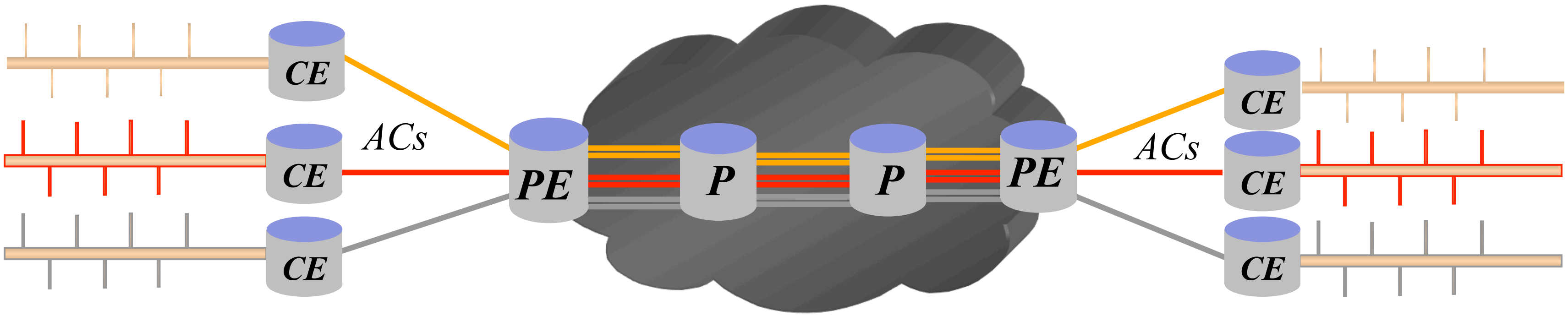


- Assume Customers 1 and 2 use overlapping IP addresses

=> then C-routers may have inconsistent tables

- Ingress PE-router pushes a label
- P-routers see only MPLS label
- P-routers don't see IP addresses - no ambiguity
- P-routers see only the MPLS label - not LAN IP addresses
- PE routers know how to map CE LANs

Naive use of MPLS for LAN Extension



Each LAN mapped to pair of (unidirectional) LSPs

Support all Layer 3 traffic types (CE is Ethernet Switch, not IP router)

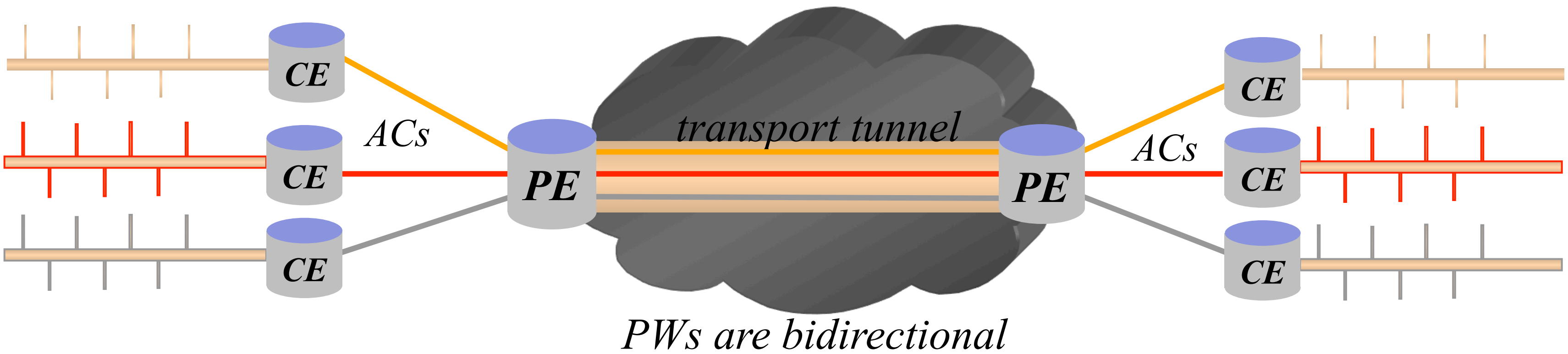
Each Ethernet frame encapsulated with MPLS label

Support various Attachment Circuit (AC) technologies

Scaling problem:

- requires large number of LSPs
- P-routers need to reserve resources for each LAN instance

(Martini) Pseudo Wires (PW) RFCs4447,4448



- Transport MPLS tunnel set up between PEs
- Multiple PWs may be set up inside tunnel
- Ethernet frame encapsulated with 2 labels
- P-routers do not reserve resources for each VPN instance

MPLS (outer) label
PW (inner) label
Ethernet frame

More on Pseudo Wires (PWs)

- Ethernet-over-MPLS Encapsulation format defined in RFC4448, L2 can be Ethernet,
  - Conceptually, L2 can also be ATM or Frame Relay (FR)
- Setup via PW control protocol based on targeted LDP

  RFC4447

Problems:

- Support only point-to-point LAN interconnect (VPWS)
- Need to manually configure PW for every VPN instance
- Need to setup 2 unidirectional tunnels for every pair of PEs

Ethernet Pseudo Wire packet

outer label	*inner label*	*control word*	***Ethernet Frame***

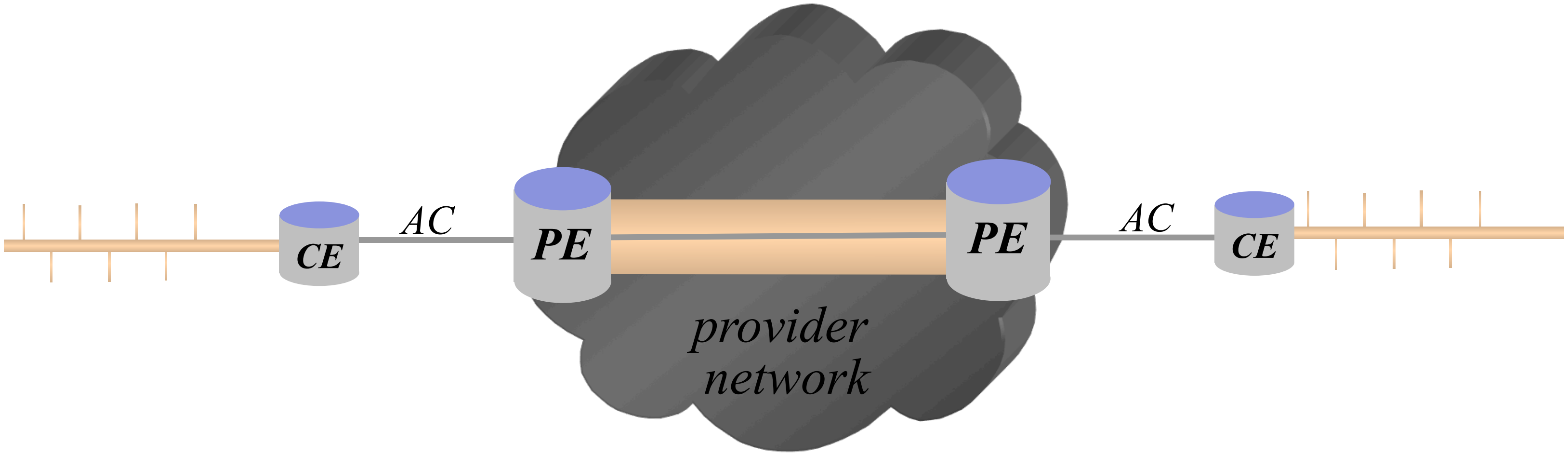
- *outer label specifies MPLS tunnel*
- *inner label contains PW label to support multiple Ethernet PWs in a single MPLS tunnel*
- *optional control word*
  - *enables detection of out-of-order and lost packets*

0000	*reserved*	*Sequence Number (16b)*

- *Ethernet Frame*
  - *by default no FCS trailer (but there is separate "FCS retention" draft)*

MPLS L2VPNs

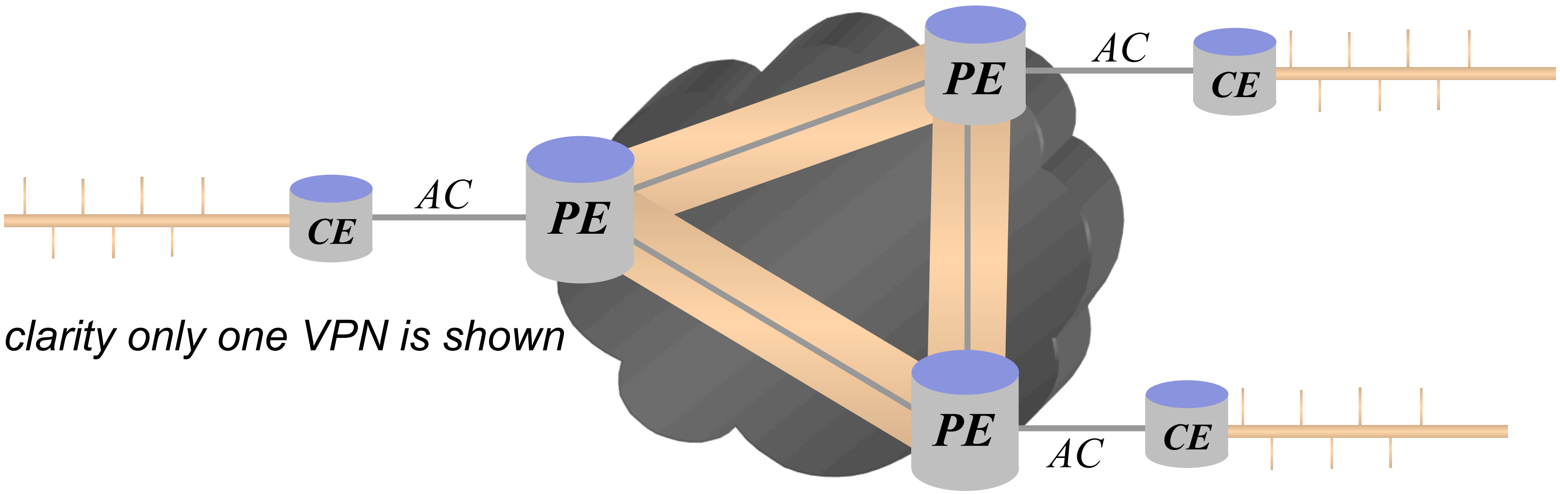
VPWS



- Virtual Private Wire Service is a L2 point-to-point service
- It emulates a *wire* supporting the Ethernet physical layer
- Set up MPLS tunnel between PEs
- Set up Ethernet PW inside tunnel
- CEs appear to be connected by a single L2 circuit

(can also make VPWS for ATM, FR, etc.)

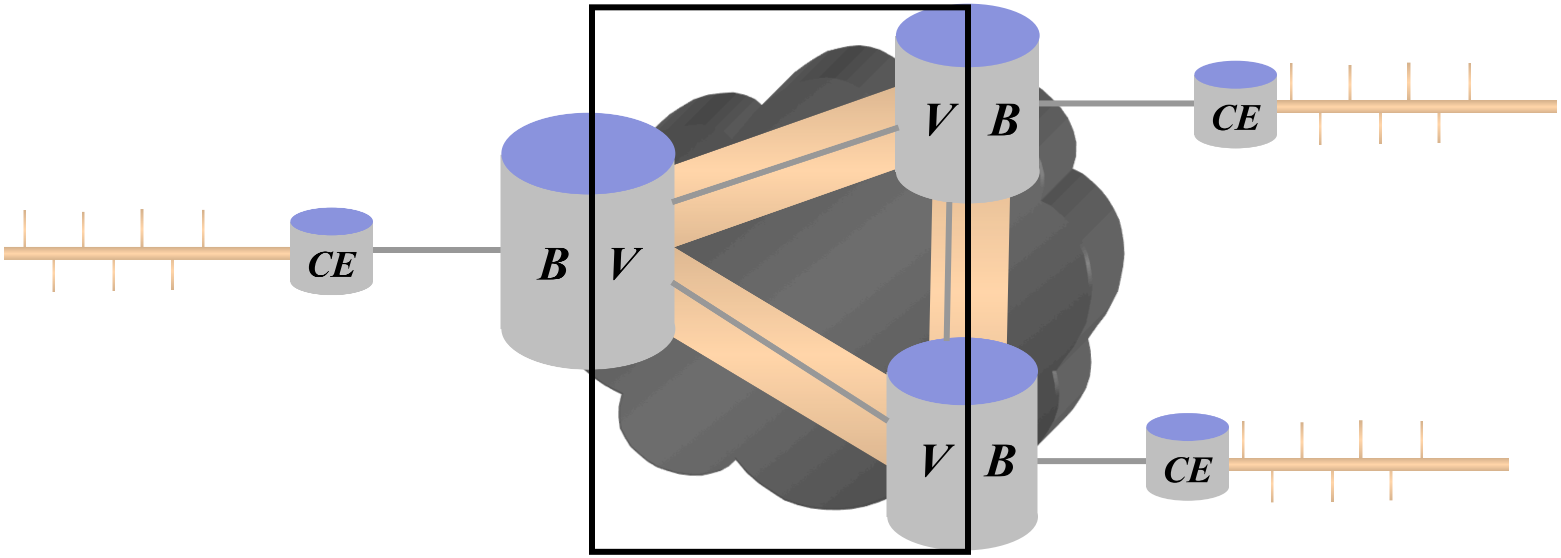
Virtual Private LAN Service (VPLS)



for clarity only one VPN is shown

- VPLS emulates a LAN over an MPLS network
- Set up MPLS tunnel between every pair of PEs (full mesh)
- Set up Ethernet PWs inside tunnels, for each VPN instance
- CEs appear to be connected by a single LAN
- PE must know where to send Ethernet frames …
  - but this is what an Ethernet bridge does

VPLS (RFC4664)



A VPLS-enabled PE has, in addition to its MPLS functions:

- VPLS code module (IETF RFC 4761, 4762 for L2VPN/PE discovery/ configurations)
- Bridging module (standard IEEE 802.1D learning bridge)
- The Service Provider (SP) network (inside rectangle) looks like a single Ethernet bridge!
  - Note: if CE is a router, then PE only sees 1 MAC per customer location

VPLS bridge module

PE maintains a separate bridging module for each VPN (VPLS instance)

VPLS bridging module must perform:

- MAC learning
- MAC aging
- flooding of unknown MAC frames
- replication (for unknown/multicast/broadcast frames)

Unlike standard L2 bridges, **S**panning **T**ree **P**rotocol is NOT used due to

- limited traffic engineering capabilities
- scalability limitations
- slow convergence

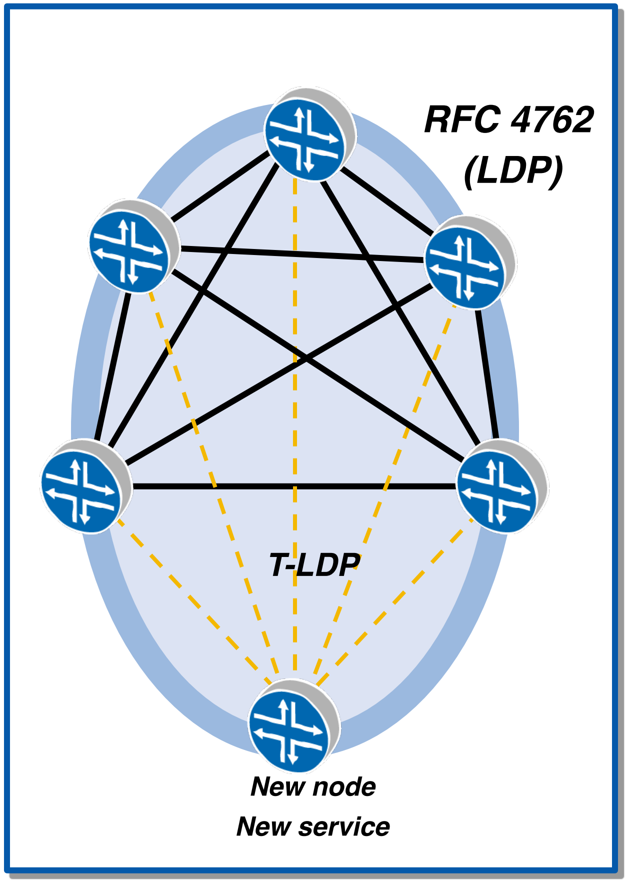
Forwarding loops are avoided by **Split-horizon**

- A PE never forwards packet from MPLS network to another PE
- REQUIRE there is a full mesh of PWs between the each PE serving a site of a VPN so that the data can always send directly to the right PE

VPLS code module

- VPLS signaling
  - establish PWs between PEs per VPLS
- VPLS auto-discovery
  - locates PEs participating in VPLS instance
- Obtain frame from bridge
  - encapsulate Ethernet frames
  - and inject packet into PW
- Retrieve packet from PW
  - removes PW encapsulation
  - and forward Ethernet frame to bridge

VPLS 2 Deployed Standards

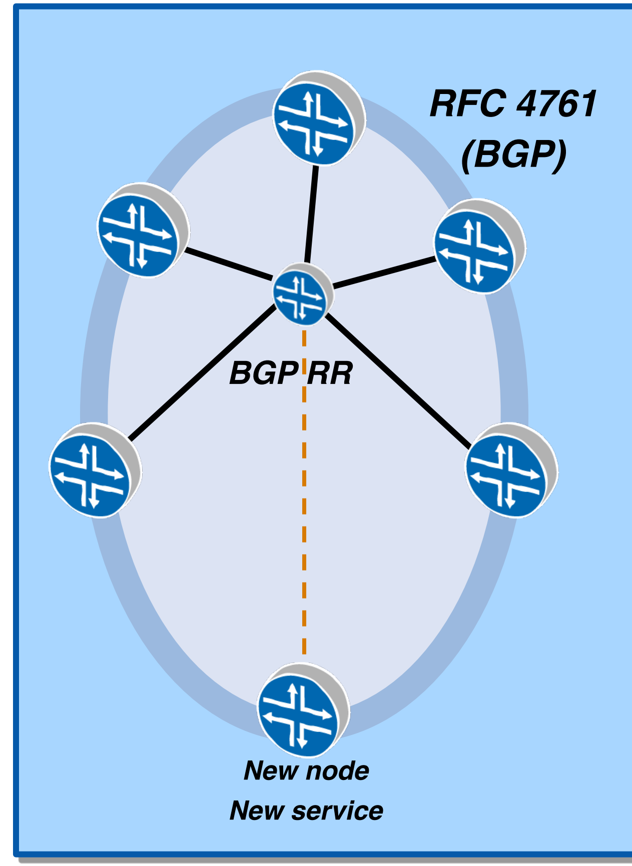


LDP-based

- ***Signaling only, no auto-discovery***
- ***High-touch provisioning***

BGP-based

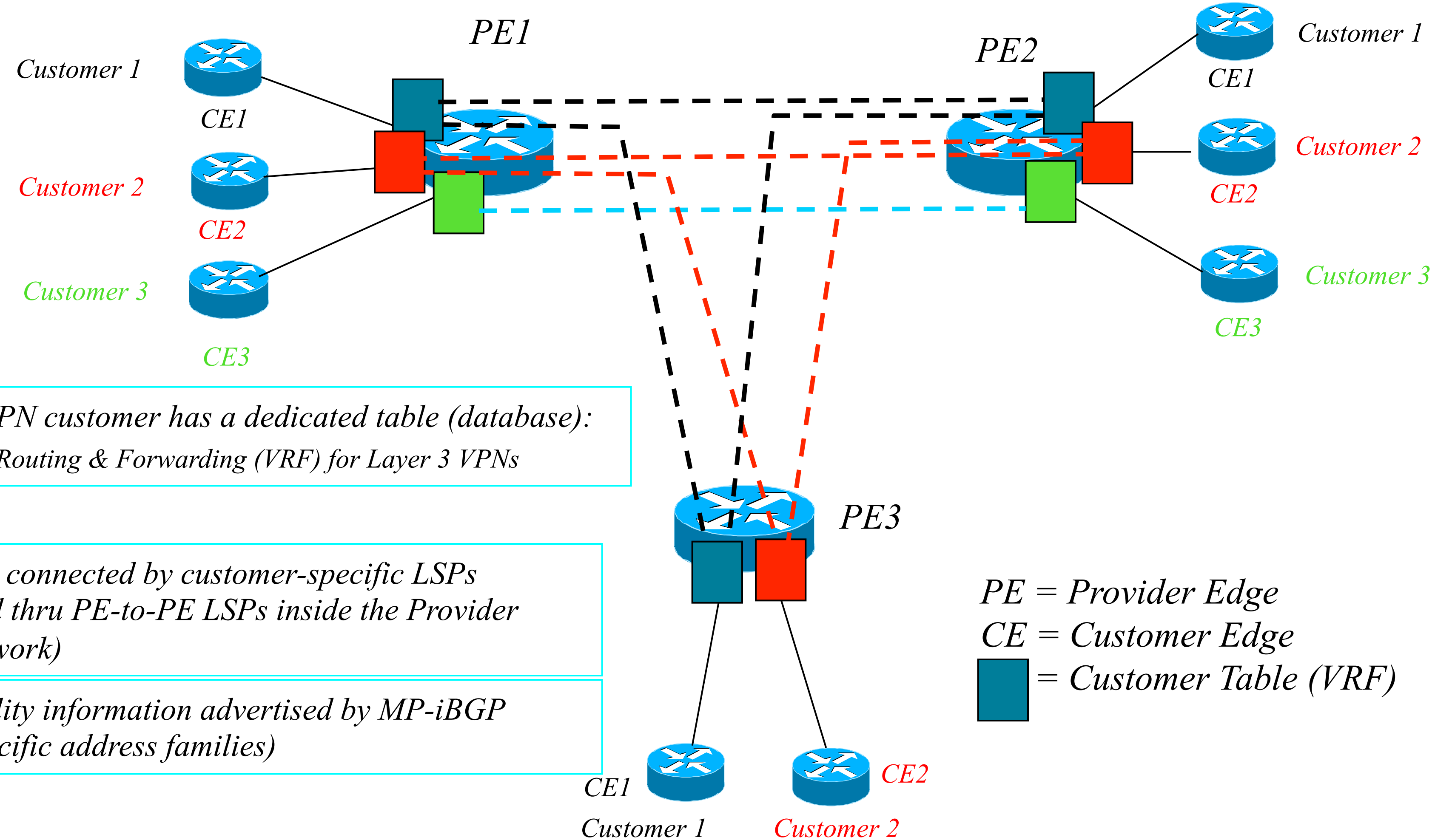
- ***Signaling & Auto-discovery***
- ***Inter-area/ metro/ provider***
- ***Multicast optimization***



Existing control-plane session
New control-plane session

MPLS L3VPNs

Conceptual View of BGP/MPLS (L3) VPNs (RFC 4364)



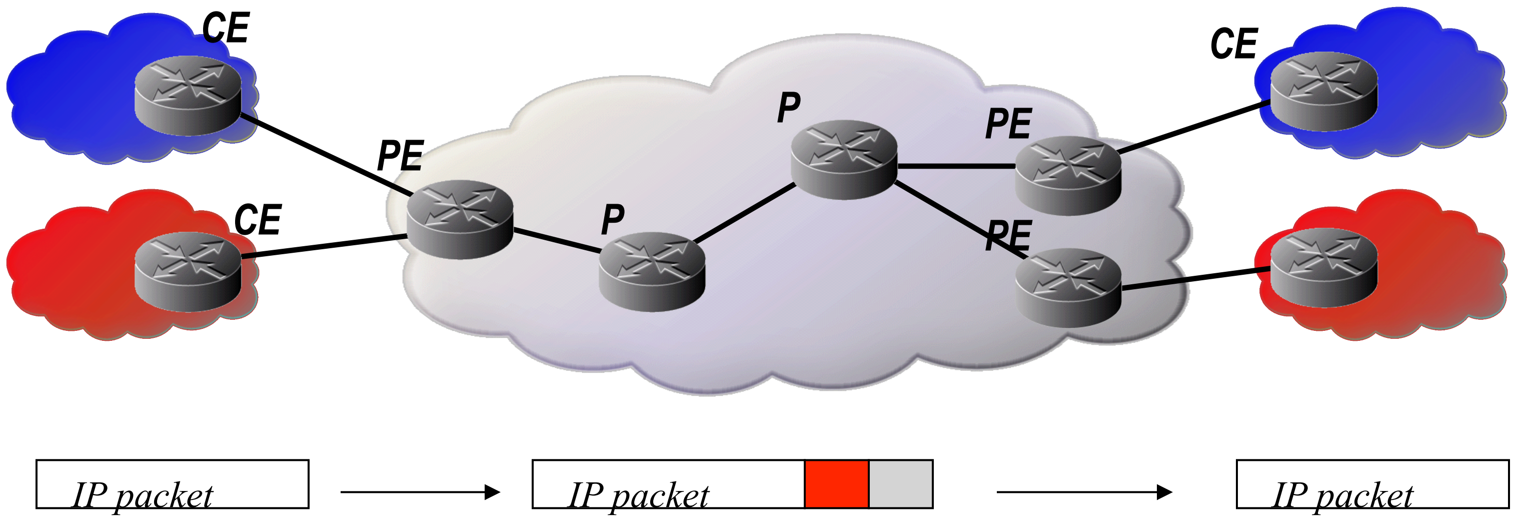
Each VPN customer has a dedicated table (database):
- *VPN Routing & Forwarding (VRF) for Layer 3 VPNs*

VPN sites connected by customer-specific LSPs (tunnelled thru PE-to-PE LSPs inside the Provider Core Network)

Reachability information advertised by MP-iBGP (VPN-specific address families)

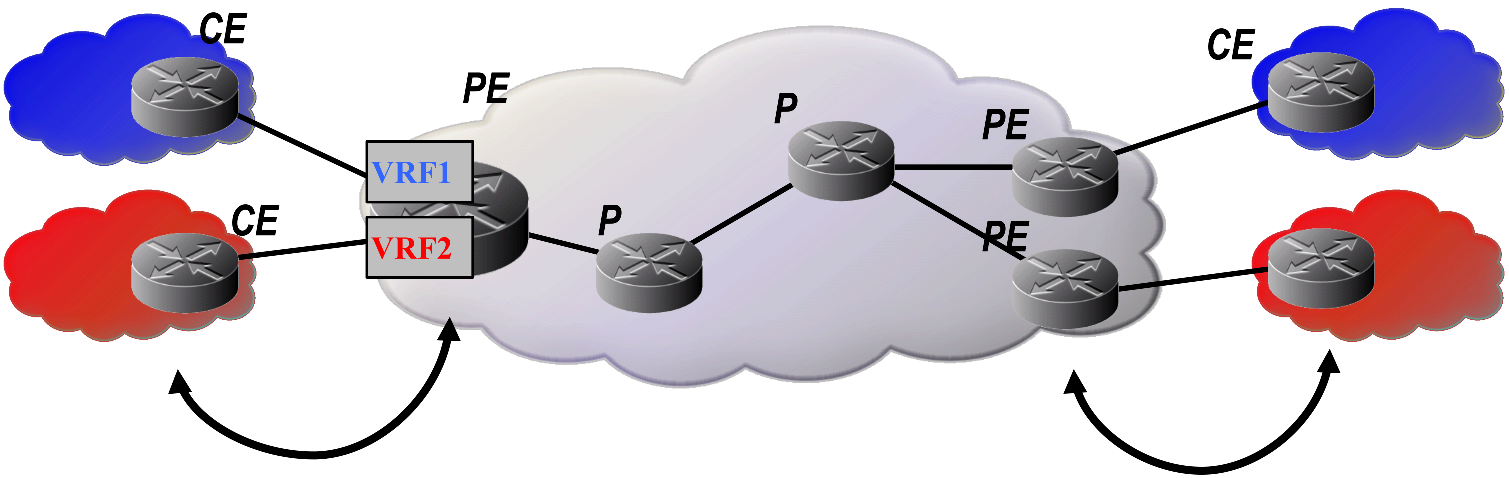
PE = Provider Edge
CE = Customer Edge
[teal box] = Customer Table (VRF)

Packet Forwarding in an MPLS L3VPN



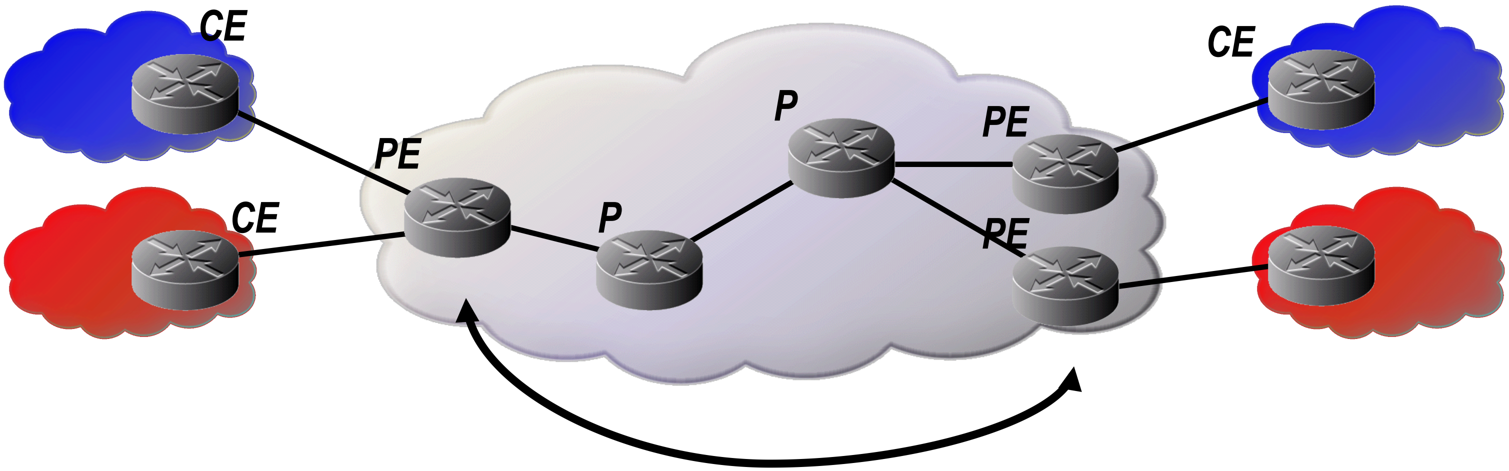
- Ingress PE router receives IP packet/Frame from CE
- Ingress PE router does IP lookup and adds label stack
- P router switches the packet/frame based on the top label (gray)
- Egress PE router removes the top label
- Egress PE router uses bottom label (red) to select VPN
- Egress PE removes bottom label and forwards IP packet/ frame to CE

PE – CE Routing Connections



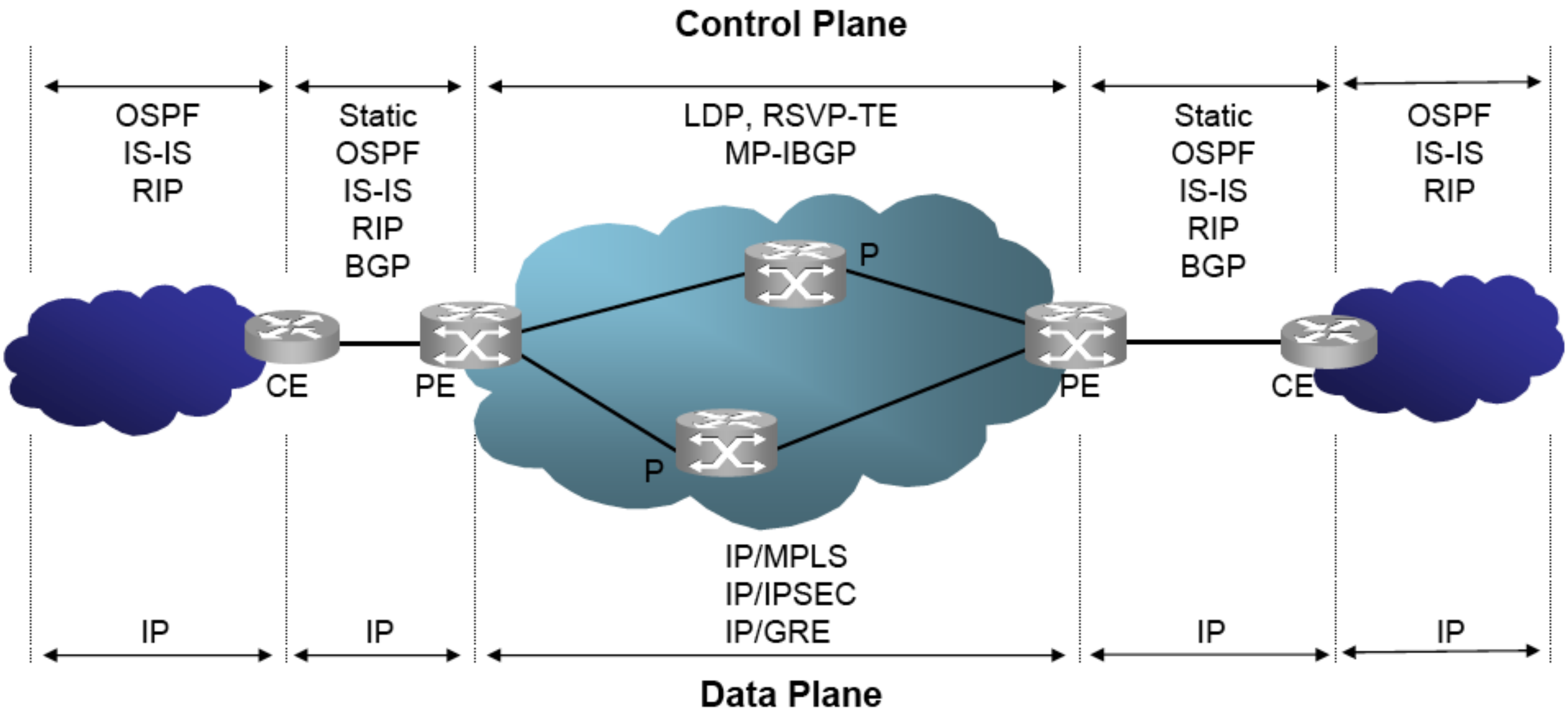
- VPN Routing & Forwarding instance (VRF) for each VPN on each PE
  - Flexible addressing
    - Support overlapping IP addresses and private IP address space
  - Secure
    - Customer packets are only placed in customers VPN
  - Customers can use different IGP; Static, RIP, OSPF or BGP
    - Each VRF contains customer routes

PE – PE Routing Connections



- MP-iBGP used between PE's to distribute VPN routing information.
  - PE routers are full mesh MP-iBGP
  - Multiprotocol Extensions of BGP propagate VPN-IPv4 routes
- PE and P routers run IGP and label distribution protocol
- P routers are VPN unaware

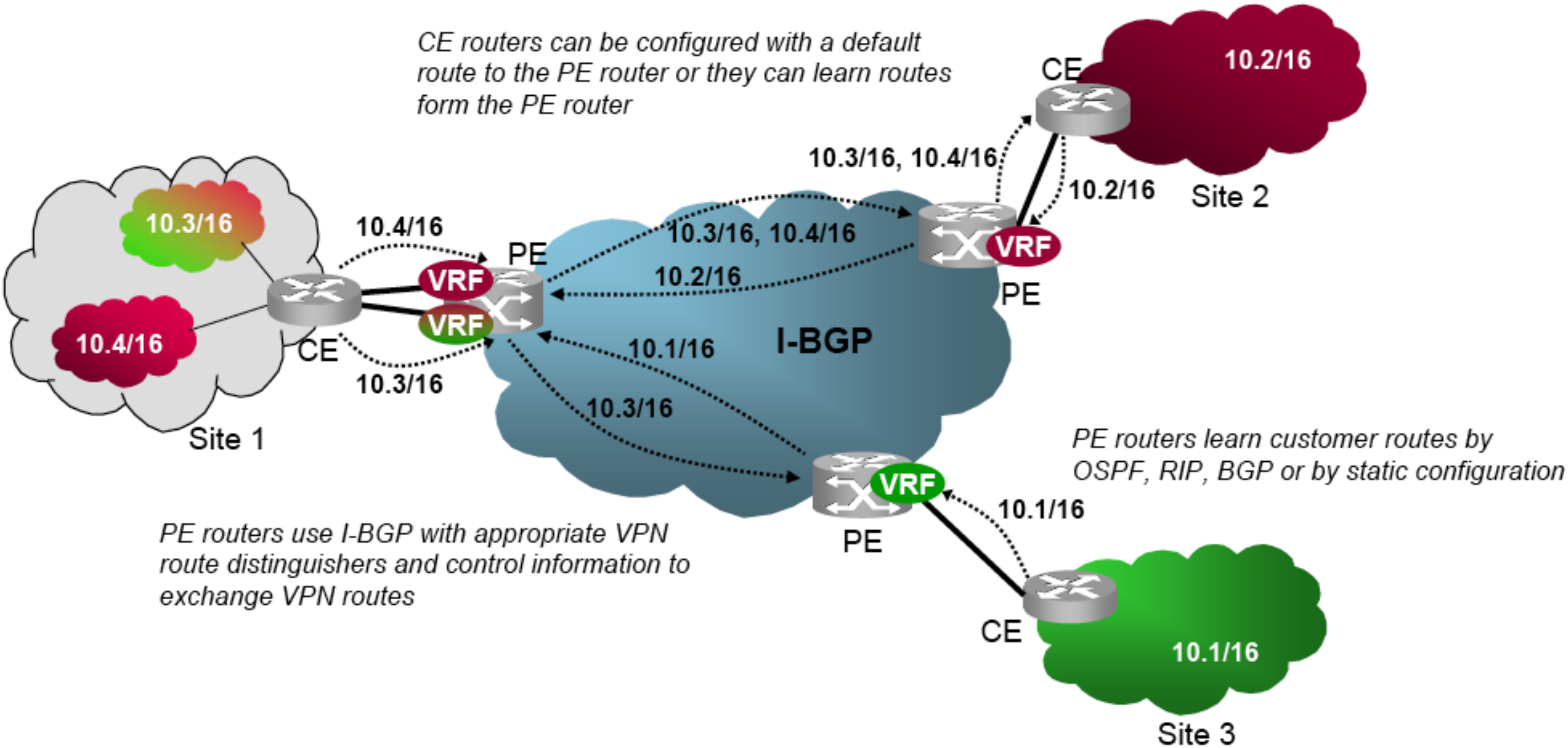
Protocols for BGP/MPLS (L3) VPNs



More Details on

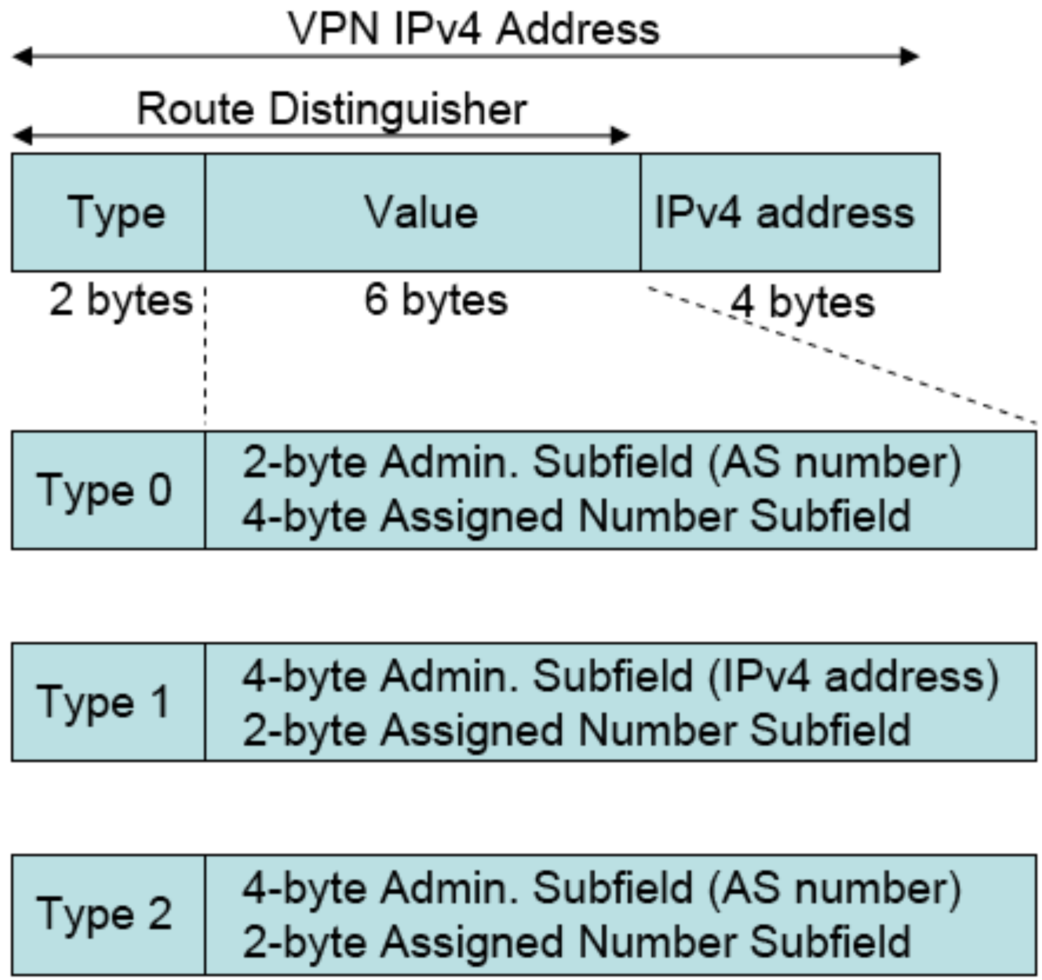
BGP/MPLS L3VPNs

Distributing VPN-specific IP addresses via i-BGP MP-extensions (RFC4364)

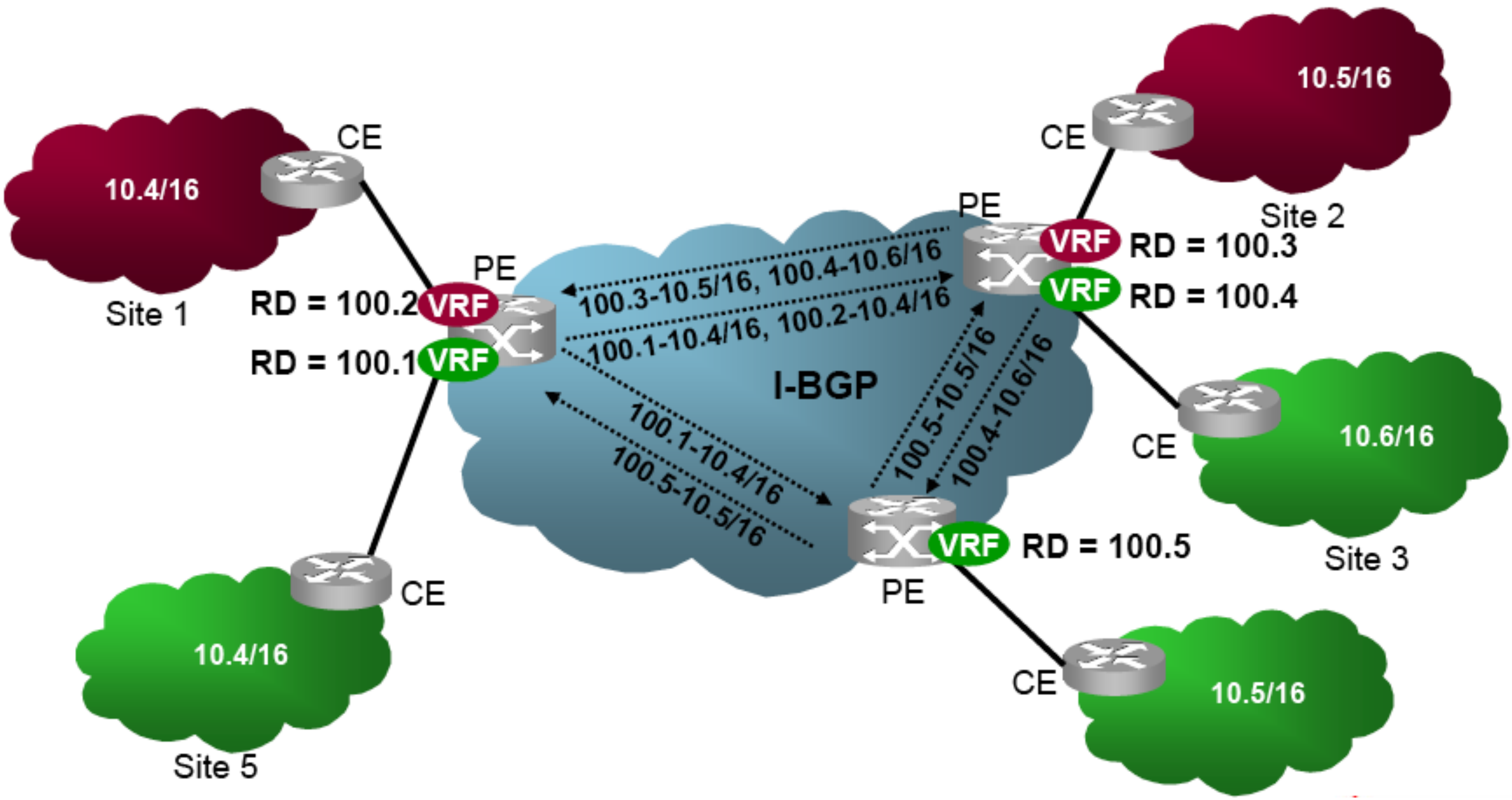


VPN-IPv4 Address Families

- BGP could not carry identical (overlapping) private IP addresses from different VPNs
- A 8-byte Route Distinguisher (RD) is introduced for this purpose
- RFC4364 defines multi-protocol extensions to let BGP to carry new type of addresses (those with RD)
- A PE needs to be configured to associate routes that lead to a particular CE with one or more RDs



Using Route Distinguisher (RD) to handle overlapping Private IP addresses from different VPNs



Route Target (RT) – A new BGP Extended Community attribute

- Key Idea: Decouple VPN-address identification (RD) from Distribution Policy (RT) to provide more VPN configuration flexibility and enhance BGP scalability
- Instead of solely using Route-Distinguisher (RD) to control the selective distribution of VPN-routes to different PEs (sites/ VRFs), an additional new BGP attribute, Route Target (RT), is defined for the such purpose
  - A route originated by a VPN-site with Export RT = "X" gets installed in any VRF within an Import RT = "X"

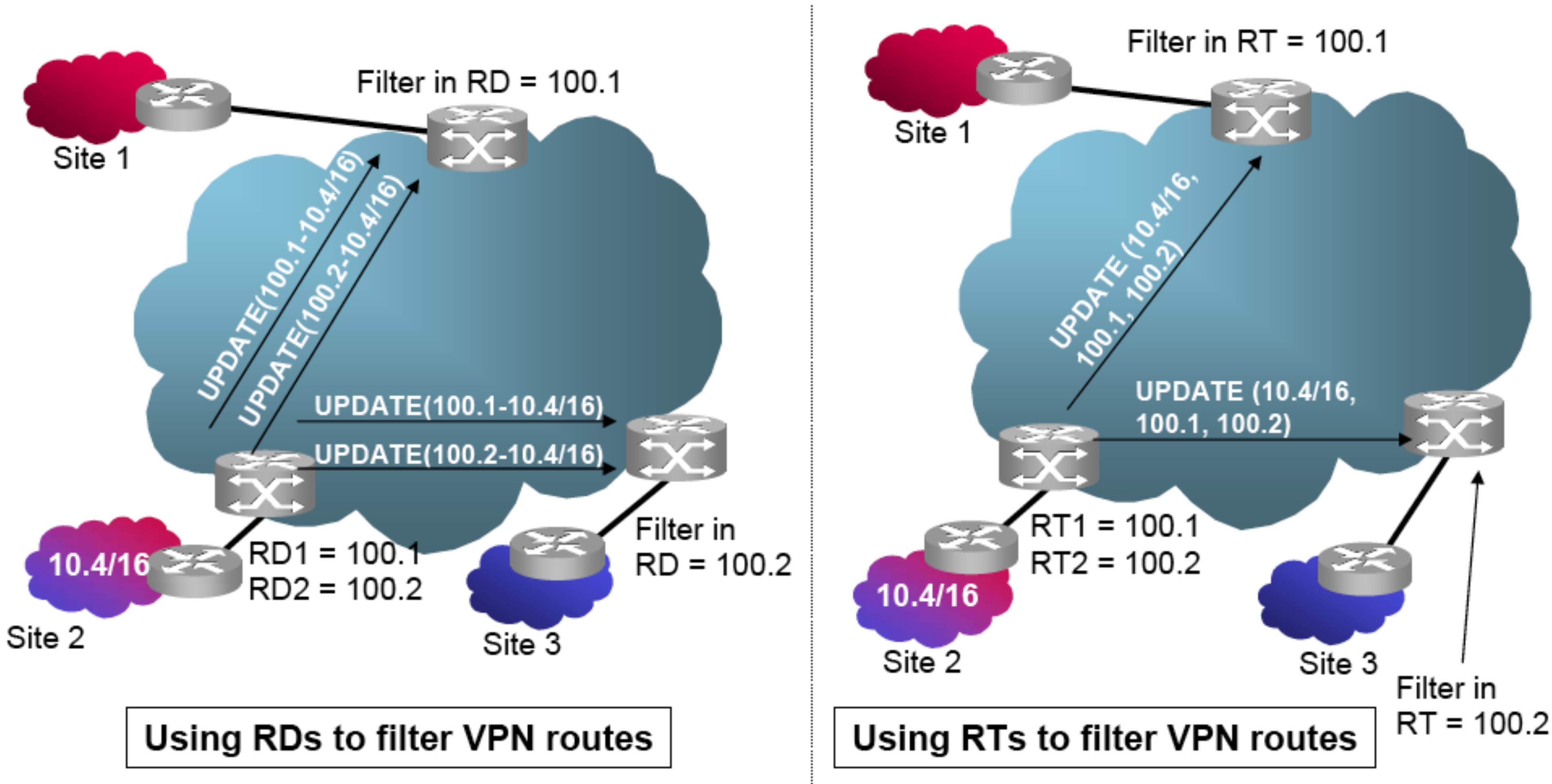
    => RT(s) are attributes of each VPN route that control which site(s) can access/use this route

    - RTs are carried in iBGP-MP as Extended Community and structured similar to the RDs
  - An alternative design could have used RDs solely to determine VPN membership of each site

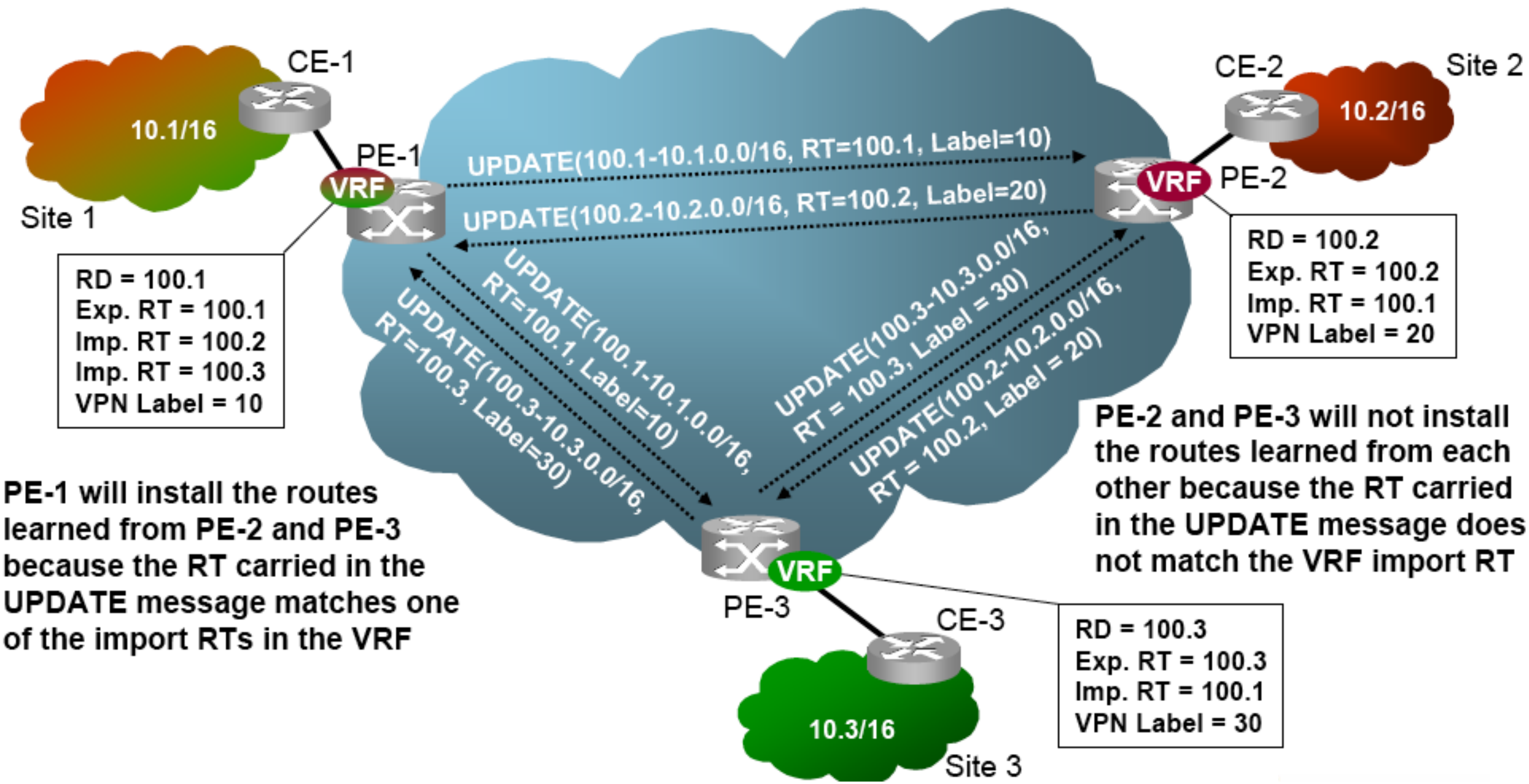
    => When a site is in multiple VPNs, its routes would need to be advertised multiple times, each with a different RD

    => Not as scalable/flexible as the current RD & RT approach

Using
Route Distinguishers (RD) vs. Route Targets (RT) to configure Selective Route Distribution/Filtering



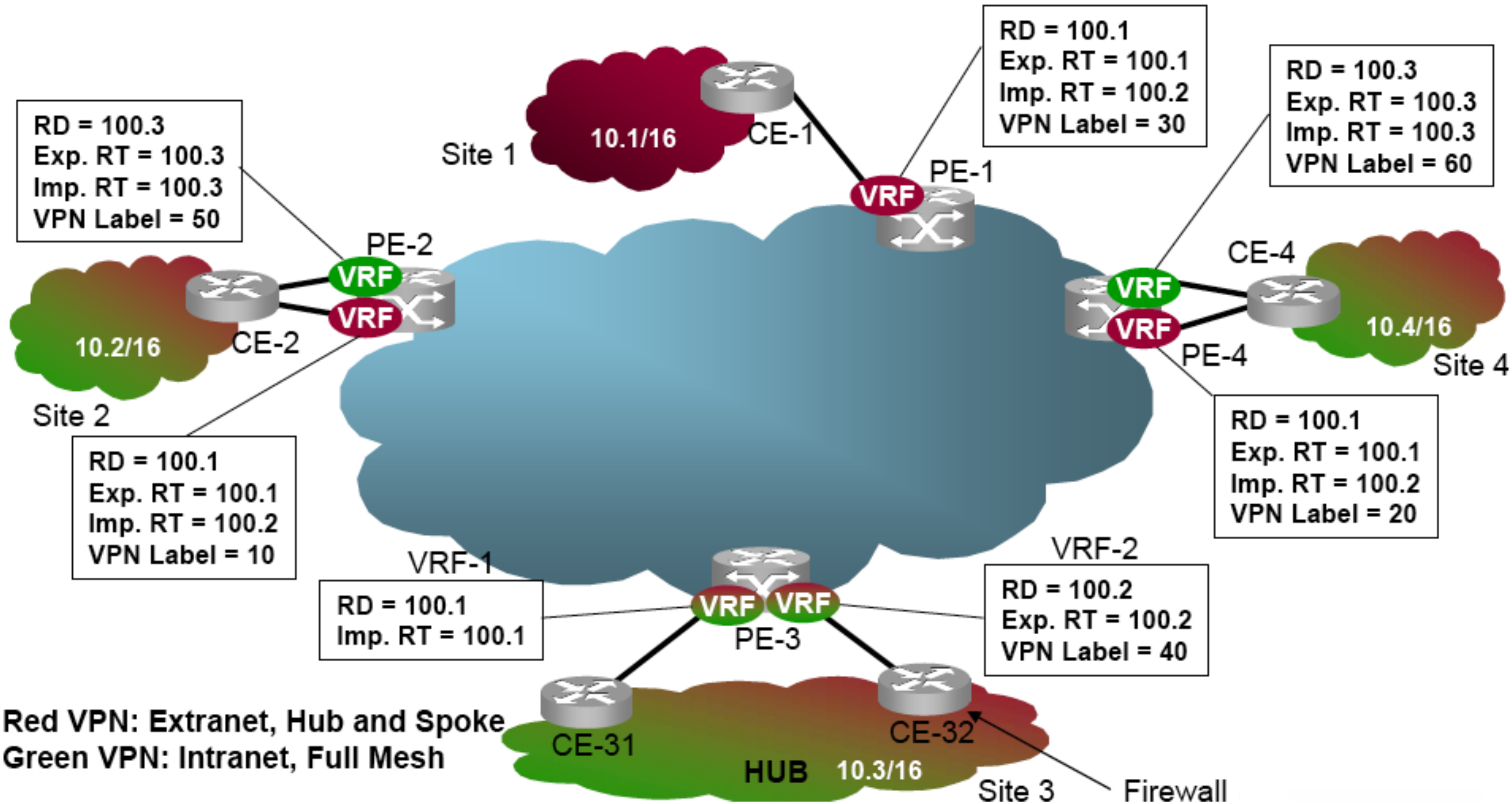
Using RDs and RTs together to efficiently support/configure Sites with Multiple VPN Memberships



PE-1 will install the routes learned from PE-2 and PE-3 because the RT carried in the UPDATE message matches one of the import RTs in the VRF

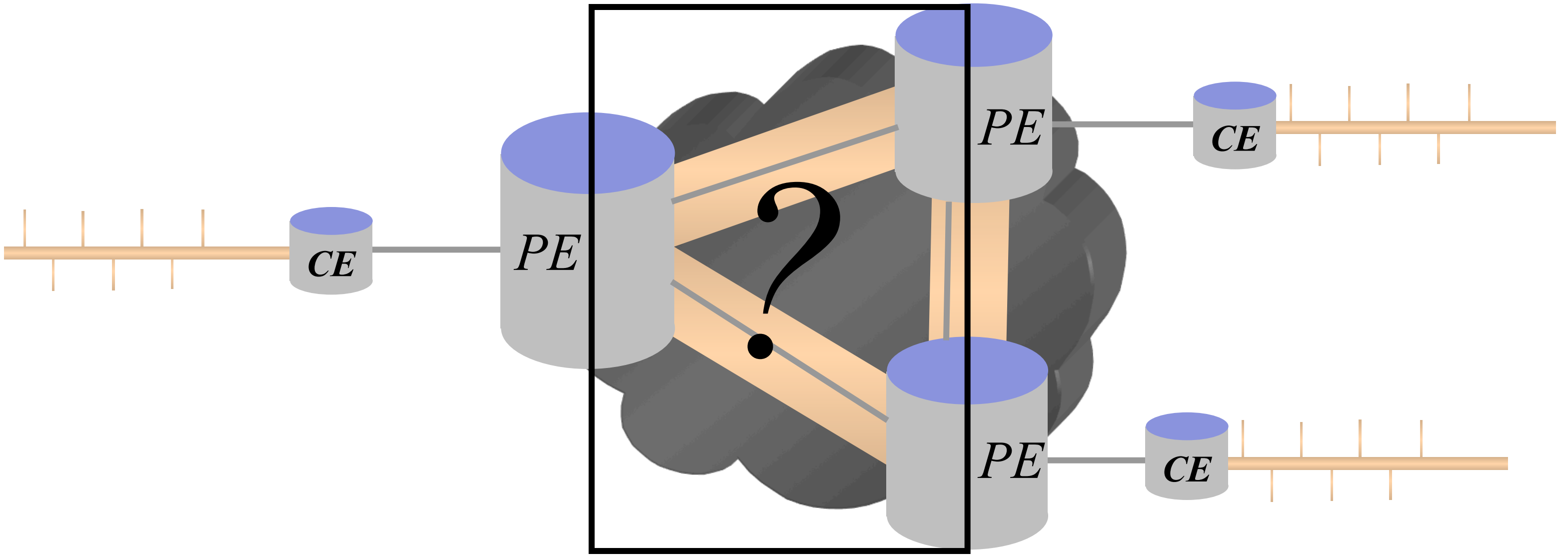
PE-2 and PE-3 will not install the routes learned from each other because the RT carried in the UPDATE message does not match the VRF import RT

Example: Support Overlapping Intranet & Extranet VPNs



Red VPN: Extranet, Hub and Spoke
Green VPN: Intranet, Full Mesh

L2VPN vs. L3VPN



- In L2VPN CEs appear to be connected by single L2 network
  - PEs are transparent to L3 routing protocols
  - CEs are routing peers
- In L3VPN CE routers appear to be connected by a single L3 network
  - CE is routing peer of PE, not remote CE
  - PE maintains routing table for each VPN

L2VPN vs. L3VPN

- C (Customer) switch connects to L2 circuits
- Signaling/Config. via BGP or LDP
- Serve all L3 traffic types of C
- Support only Ethernet as L2 tech.
- C is responsible for routing
  - "overlay model"
- Simple C-to-SP interface
- C peering scales as VPN size
  - => scaling problem

- C router peers with PE router
- Signaling/Config. via BGP
- Service limited to IP traffic
- Supports different L2 technologies
- Service Provider (SP) responsible for routing
  - "peer model"
- Complex C-to-SP interface
- C peering independent of VPN size
  - scales well

References

- Bruce Davie, Yakov Rekhter, "MPLS: Technology and Applications," published by Morgan Kaufmann, 2000.
- Bruce Davie, Adrian Farrel, "MPLS: Next Steps," Volume 1, published by Morgan Kaufmann, 2008.
- Cisco MPLS Configuration guide (Google) or
  - http://www.cisco.com/en/US/docs/ios/mpls/configuration/guide/12_4/mp_12_4_book.html